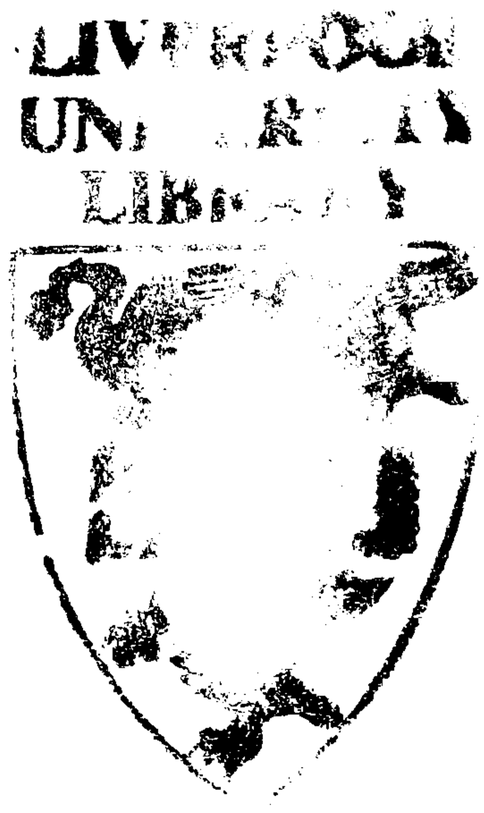
REDUCTIVE PHYSICALISM AND PHENOMENAL PROPERTIES

The Nature of the Problem

Thesis submitted in accordance with the requirements of the University of Liverpool for the degree of Doctor in Philosophy

by Brian George Crabb.

July 1996



Reductive Physicalism and Phenomenal Properties

CONTENTS

ACKNOWLEDGEMENTS

This work was researched and written during my period of part-time registration with the University of Liverpool, where my principal source of information has been the Sidney Jones Library. I also wish to express my thanks to Barry Dainton and Nicholas Nathan for their extensive and painstaking comments and suggestions. I was privileged to spend the academic year 1992 - 1993 at the departments of Philosophy and Cognitive Science, Tufts University, Medford, U.S.A., where the untiring help and enthusiasm of Daniel Dennett, Jeff McConnell and Stephen White provided me with insights into their philosophical views which would have otherwise been unavailable to me. I am also indebted to Ned Block, Norman Daniels, Owen Flanagan, Hilary Putnam and George Smith for their helpful discussions. Special thanks must go also to Harvard University and the Massachusetts Institute of Technology for their generous hospitality.

Brian Crabb July 1996.

INTRODUCTION

Introduction.

The purpose of this work is to explore the ways in which phenomenal properties, or the qualitative character of sensory experiences, might be seen to present a problem for reductive physicalism; the thesis that occurrent mental states, universals [properties of mental states] and mental events are fundamentally physical in nature. [1] The initial assumption is that for there to be a problem at all, even *prima facie*, there must at least be an apparent conflict between the claims made about qualia by the reductivist and his dualistic adversary respectively. So a major project for the thesis is to find out where this conflict occurs, and what, if anything, it amounts to. And since the project is conducted in accordance with a rather specific strategy, it is essential at the outset that the logical structure of our strategy be explained, at least in broad terms. As we shall see later, the strategy we wish to adopt is perhaps not as crisply definable as we would like, but it is at least possible to outline the logical structure we would *like* it to have.

Firstly, reductive physicalism must presumably be an intelligible thesis about the nature of the world; with respect to qualia and experience in particular, it must at least provide an intelligible account of what is going on when we take ourselves to be experiencing qualia. Our second initial assumption is that for qualia or experience to present even a prima facie problem for that world-view the dualist must be making some intelligible claim about those phenomena which at least *appears* to conflict in some significant way with the physicalist's thesis.

---

1. There is at least one compelling reason for taking this thesis as our starting point; namely that if such an identity relation were found to obtain, the problem of how mental and physical phenomena interact causally, if at all, would no longer be a problem. A phenomenon does not causally interact with itself.

Having taken on the first two assumptions as a working hypothesis, then, the strategy will be to try to infer what the prima facie problem is from what the dualist says about qualia and experience in his various attempts to rebut reductive physicalism. More specifically, given that each of his various arguments has a certain characteristic structure, we shall be asking what he *needs* to say, within the framework of each argument, about qualia in order to present a problem.

To this end, then, there are three quite distinct questions to be addressed, and for brevity I shall draft these questions specifically in terms of *qualia*. The questions we would ask about experience *per se* might then be drafted in parallel fashion.

Firstly, then, what *is* the essential thesis of the reductive physicalist, RP?

The second question might have been simply "how does RP propose to accommodate occurrent qualia?", but here we would run into an immediate conceptual difficulty. For although we might have a quite vivid *epistemically* based notion that there are, or at least seem to be, qualia we can hardly claim to be able to provide a definitive account of their metaphysical status, or of the properties which essentially belong to them, without further argument. If we could do that, it might be possible to see straight off whether or not they can be accommodated within any of the various physicalistic accounts of the world. What we shall find is that each of the proposed qualia-based counterarguments to reductive physicalism depends for its force on quite distinct claims about qualia, and unless we simply presuppose that we know everything relevant about qualia at the outset it would seem more appropriate to see firstly what those claims are. The question of how the physicalist would *need* to accommodate each of the counterarguments in turn can then be approached in the light of those specific claims. So our second question will be, rather, "what facts about qualia does the dualist cite in each counterargument as a *prima facie* problem for reductive physicalism?"

Once the first two questions have been answered at least in broad terms, then, the third becomes obvious; it will be "do qualia *have* the attributes required of them by any of the proposed

counterarguments?". For only if they do might there be even a *prima facie* problem for reductive physicalism in virtue of which he should be required to refine his thesis.

The Physical.

The first commonsense assumption we shall be making is to the effect that there is a physical realm, and that all phenomena which fall within that realm are, insofar as they are at least in principle capable of being understood, observed and individuated at all, capable of being understood, observed and individuated from a third person point of view. Thus, a phenomenon will only be counted as physical if the third person point of view is capable of providing epistemic and conceptual access to all the facts about that phenomenon which can be accessed. Naturally, there are those who will strongly contest this commonsense construal of the physical, and for a variety of reasons, but unless good reason is found for supposing that the construal is either unintelligible or unacceptable, we shall take it as a minimum requirement.

Physicalism, then, will be construed somewhat autocratically as at least *entailing* that a complete account of the world must be ontically committed only to states, properties, events and so on which are, at least in principle, both cognitively and epistemically available from within a third person perspective. In the case of qualia, this amounts to the thesis that there will be no ontic commitment to qualia except insofar as such properties are so available. Now there is nothing very mysterious here, either about the claim being made or our motivation for characterising it as such. For although "physical" might be construed in some more technical sense for other reasons, our chief interest, or so I have supposed, is in whether qualia comply with our commonsense intuition as to what there is in the *objective realm*; what states, properties and events we might reasonably accept as belonging to the realm of publicly intelligible and discernible phenomena. Here, for a phenomenon to be publicly intelligible would be, roughly, for it to be possible to provide a complete descriptive account of that phenomenon without recourse to any concepts or predicates which cannot be rendered *interpersonally* intelligible. Just as we are able to understand one another's talk of tables, atoms and the so-called

"primary qualities", then, we should also expect to understand our talk of particular qualia, construed as properties of or in experience. There is no doubt that the concepts invoked by that expectation will be worthy of further refinement, but we might reasonably assume that we have at least provided an intuitively intelligible thesis as our starting point. Similarly, we might at least assume that it makes good common sense to speak of such phenomena being *epistemically available* in the third person perspective. For to the extent that you might reasonably expect to be able to corroborate or refute my claims about tables, atoms and the space-time continuum, at least in broad terms, you should also expect to be able to corroborate or refute my account of the particular experiential quality I claim to be typically associated with my perception of red objects.

Reduction.

The second conceptual nexus we need to consider is that invoked by the thesis [QR] that qualia are *reducible* to physical properties. Again, there seems to be no compelling reason to avoid a commonsense interpretation of this claim at the outset. Thus, while it is common to appeal for clarification to such notions as "supervenience" or "constitution", I see no reason not to say that for [QR] to be true, qualia must simply *be* physical properties. For just as we are entitled to assume within the customary physicalistic framework that talk of tables is only reducible to talk of physical objects if tables *are* physical objects, we might have a similar expectation of qualia-discourse. A quale will be counted as reducible to a physical property just if it *is* a physical property. So if an occurrent quale being constituted by a physical property does not entail its *being* a physical property, a reductive programme framed in terms of constitution will have failed. We should note, however, that there is a certain degree of flexibility attached to this requirement. For even if the set [S] of paradigmatically physical, or physico-dispositional [PPD] items already recognise by science does not include qualia, it remains possible in principle that they should nevertheless comply with the criteria for *annexation* on to the prevailing ontology. Clearly, whether they can be added on will depend at least in part on what those criteria turn out to be. In this respect then, we might even deny that there is an intelligible

distinction between reductive and non-reductive physicalism. For while reductive physicalism requires that qualia should be members of [S], we might suppose that there is a sense in which the content of [S] has yet to be established.

The First Person Perspective.

Another assumption to which we shall initially take even the reductive physicalist to be committed is that there is also a *first* person point of view with regard to one's own occurrent states, properties and events. If I stub my toe on the step, I am capable of noticing that something has happened to me without having to resort to the third person perspective. For while I might notice that the event has occurred by observing my own responses and physical state in the third person perspective [I might notice that I am hopping about on one leg and clutching my toe, catch a glimpse of myself in a mirror and notice that I am wincing visibly or even observe that my toe is bruised and swollen], I can nevertheless notice that *something* has occurred without having to make any of these observations. Typically, I would claim to notice that I experience a sudden pain in my toe. Whatever it is that I thereby notice, it seems an undeniable matter of common sense to affirm that I notice in the first person perspective that *something* has happened to me.

States, Properties and Events.

Common sense will also be applied initially to certain other concepts invoked by our project. Thus, the conceptual relation between states, properties of states and events will be construed in the following way. When I stub my toe and notice that something has happened to me, I might initially describe that state of affairs in first person terms by saying that an event e has occurred at time t. I might then offer the further observation that the event e which has occurred amounts to a change at time t from being in one state to being in another. Thus, before I stubbed my toe, I was in one state [a state in which I do not notice that the particular something has occurred], while upon stubbing my toe I enter another state, in which I do notice that the particular something has occurred. To that extent, then, I am committed to the further claim

that the consecutive states are of different *types*. There is some property or other which is possessed by the latter state but not the former [or vice versa]. And *to the extent* that I am able to characterise those distinct types of state from the first person point of view, reductive physicalism ordains that I am able to do likewise from the third person point of view. Hence, in brief, it is committed to the claim that insofar as I am able to notice in the first person perspective that a particular event of an identifiable type has occurred at time t, there is nothing in principle to prevent someone else understanding and noticing all of this *as such* in the third person perspective.[2] Naturally, if this brand of physicalism is to be regarded as complete, the same must be said of the process or state of noticing, or being aware, of the facts of which I *am* aware in introspection.

Types and Tokens.

Now it is customary to draw a fairly firm conceptual distinction between *token* occurrences and token occurrences of a particular *type*. Thus, while common sense dictates that I am able to affirm in the first person [introspective] perspective that my state has changed from one type to another, it appears to be quite another matter to provide an account of the particular types of state involved; an account of the properties [construed as universals] belonging to each state. I might say simply that in the first place I noticed that I was in a state of the type [Not being in pain], and that in the second I was in a state of the type [Being in pain]. But there is no presumption here that my characterisation of the two distinct state types provides a complete account of each state. For, in parallel fashion, I might describe two distinct physical objects as large and small respectively, even though I have provided no

---

2. The thesis here is not just that Smith, say, will be able to notice [or infer] that Jones is in a particular kind of *mental* state, but that the mental state will be intelligible, if at all, within the conceptual framework of an essentially third person account of the world.

other information about the objects involved. Thus, it might be stipulated that an object is of the type [Large] just if it has spatial dimensions in excess of a particular quantity, and of the type [Small] if it does not. Being of the type [Large] is then simply having the *property* of having certain minimum dimensions. Similarly, then, if a particular event e is discerned at time t in introspection, we must at least concede that a state of some type or other occurs, or comes into being, at time t; that I enter a state which has some property or other which was not possessed by the state which preceded it. The question of exactly which property that is might remain unanswered even in principle.[3] All we need say is that if reductive physicalism is true, then insofar as information concerning that property is available in introspection, it is also available in principle in the third person perspective and

---

3. This construal of properties as *universals* which determine which *type* any particular token item [e.g.,state or experience occurring at time t, etc.] belongs to will be adopted throughout our discussion. It is, however, relative in the following sense. Firstly, all token items which have a particular property [e.g., 'Redness, or R] will be said to be of the type [Red], or [R], even though the items of this type need have no other properties in common. So, for example, a ripe tomato and a sample of blood will be of the same type [Red] just in virtue of sharing the property *Redness*. On the other hand, the two items of this type will each have properties which the other does not. Ripe tomatoes are of the type [Fruit], just in virtue of having the property of being a fruit, while samples of blood are not. Whether or not distinct token items are of a single type will therefore depend on which property we are considering. Secondly, it is possible to build up a conceptual hierarchy of properties, and hence of types. Thus, at the next level, there are distinct *types* of Redness. Items which have the property of *"Scarletness"* and items which have the property of *"Crimsonness"* are all of the type [Red] in virtue of sharing the property of *Redness*. Nevertheless, we might legitimately regard Crimsonness and Scarletness as being different properties, and therefore construe red items which are scarlet as being of a *different* type from red items which are crimson. When we speak of token items as being of a particular type, therefore, we are saying

intelligible within the framework of a third person conceptual scheme. More generally, insofar as any information is available and intelligible in introspection, it is also available and intelligible in principle in the third person perspective.

We can now consider in what sense if any our reductive physicalism is committed to a type-type correlation between the mental and the physical. Suppose that I am able to determine in introspection just that at time t the state I enter has properties x, y and z. I can then affirm that in introspection I can discern that an event e occurs at time t, such that the occurrence of e at time t amounts just to my entering a token state which has properties x, y and z; in other words, that I enter a state of the *type* [xyz] at time t. Reductive physicalism then requires that this fact can be understandable and determinable at least in principle in the third person perspective. Hence, if the latter is possible, someone else is able to determine [intelligibly] in the third person perspective that I enter a state of the type [xyz] at time t. The particular state referred to is a *token* state of the *type* [xyz] and to that extent properties x, y and z must be intelligible and discernible in the third person perspective. It follows, therefore, that the properties x, y and z which are intelligible and discernible in the third person are identical with the properties x, y and z discernible and intelligible in introspection, and to that extent my thesis is a type-type identity thesis. All introspectible [mental] events are occurrences of states whose introspectible properties are physical properties.

---

just that there is a particular property which they share. We might then be required to explain which particular property that is. Following on from this, we shall understand token items of a particular type, for example [Red], as manifesting particular *occurrences of Redness*. We shall then adopt a colloquial convention for referring to occurrent *tokens* of the property Redness. If the property of *Phenomenal Redness* is Rp, then, if Smith experiences a token of Rp on a particular occasion we shall say simply that he *experiences Rp* on that occasion.

What we also presuppose, however, is that different occurrences of an *introspectible* type, [Pain], for example, must be of the same physical type. For while it seems intuitively that my numerous headaches might have no single physical property in common; that there is no prima facie reason to suppose that the property H borne by each headache is of a single *physical* type, our reductive physicalism dictates to the contrary. For if it is true that each headache has a single introspectible property H discernible in introspection, that single property H must be intelligible to the physicalist in the third person perspective. He must be able to understand in principle what it is that each pain has in common; otherwise, there is some fact which is intelligible in introspection but not within the third person conceptual framework. It is perfectly plausible to suppose that a number of quite distinct physical state types [i.e., state tokens which do not share *all* the same properties] each count as headaches in the first person perspective, but for the physicalist they must at least have a single property in common which is intelligible and discernible in the third person perspective, and in virtue of which he can understand what makes each of them count as being a pain. If the reductivist's thesis is true, that property must be property H.

This is to be contrasted with a weaker token-token identity form of the thesis, which holds more modestly that although each and every mental state is a physical state, distinct token occurrences of that mental state might nevertheless share no *single* property which is intelligible and discernible in the third person. So for example, token-identity theorists of this ilk might claim that when I am in pain, the state I am in is purely physical, but that some of the particular properties of that state which are intelligible and discernible in introspection are not particular physical properties. On this account, a pain discerned in introspection as having property H can be discerned in the third person as occurring, but distinct occurrences of H might be occurrences of either physical property x, y or z, such that each of these properties has nothing in common in virtue of which a state bearing it is to be counted as a headache. If this latter position can be referred to as physicalism at all, it is not *reductive* physicalism in the strict sense; for it entails that there are introspectible and discernible properties which are not physical properties. Although this might turn out to be the only plausible option open to the physicalist, it

is not the position we set out initially to evaluate. In order to comply with our physicalistic expectations it would be necessary to deny that [H] is a single introspectible type after all. More generally, it would be necessary to insist that there are no single introspectible types which are not single physical types.

Topic-Neutrality and A Posteriori Physicalism.

If, according to my version of reductive physicalism, the introspectible properties of all occurrent states are physical properties, it might appear that I am committed to a form of *analytic* reductivism. For if each token of an introspectible and intelligible property x is identical with some intelligible physical property X, it seems that the properties I report in introspection can be construed as being *conceptually* identical with physical properties. But we must be careful here. For while it would be true to say that my reductive physicalist will be able to understand all introspectible properties *as such*, it does not follow from this that what he *means* when he reports an occurrent property x in introspection is that there is an occurrence of *physical* property X. It is epistemically plausible to suppose that the introspected property x is recognised by him only as such, and that it is a matter for a posteriori investigation to discover whether x is a particular physical property X. As we shall observe later on, it is possible for a subject to be in an epistemic situation in which he can discern that he has a headache [i.e., discern that he is in a state which has property H], but not also be able to discern in that situation that property H is a particular physical property. In order to establish this latter fact he would need to have a complete grasp of the relevant physical concepts *as such*. Hence, he might be in a position to determine in introspection that his pain is a state of type H, and property H might be a physical property, but he might nevertheless lack a sufficient command of the facts in the third person perspective to determine that H is a *physical* property at all. Consequently, we must concede that our position is a type-type identity thesis, insofar as any state of type H which is discernible as such in introspection is a physical state of a particular type, but not that having the [introspective] concept of H entails having the concept of that state *as a physical state* of a particular type. Thus, the fact that a person has introspective access to events

which are of physical type [RP][4], and which he picks out introspectively as being experiences of type [Rp], does not even entail that that he understands property Rp in the third person perspective. Hence, it would be implausible to suppose that his *concept* of what are, *in fact*, experiences of RP [see footnote 3], is identical in content with his concept of events of type [RP].

In summary, then, we can say that for the reductive physicalist every state which is discerned in introspection is a physical state, and that every property which is possessed by a state discerned in in introspection is, to the extent that it is discernible and intelligible in introspection, also discernible and intelligible within the physicalist's third person conceptual framework. To the extent that I am able to discern intelligibly in introspection that I have a headache, for example, the state type I am discerning must also be discernible and intelligible in principle in the third person perspective. But there is no requirement that the information thus gained introspectively must logically or conceptually entail facts as framed in the third person physicalistic account. It might be possible for me to know that I have a headache, and that it has certain characteristics, without even realising that either the headache or its characteristics are purely physical states' or properties. I propose to adopt all of the above commonsense assumptions uncritically at the outset in order to embark on the main project; to find out what the dualist would need to say about qualia or experience in order to present even a prima facie problem for reductive physicalism. There is, of course, no guarantee that what appears, prima facie, to be true, or even intelligible, as a matter of common sense will turn out to be so. Nevertheless, my conservatism dictates that common sense will prevail unless and until it is found to be philosophically problematic.

I remarked earlier that I would have liked the project to be as simple as it first seemed, and we might gain some consolation from the evident fact that it is at least intelligible. Nevertheless, there are complications. For what the dualist will need to say will

---

4. We take an *event* to be of the type [Rp] just if it entails entering a *state* of the type [Rp].

depend on the type of anti-reductivist strategy he chooses to adopt. But this entails in turn that how the the reductive physicalist will attempt to *accommodate* the dualist's claims about qualia will also depend on the dualist's chosen strategy. Initially, then we shall take it that according to reductive physicalism the ontic commitments of *current science* are to the exhaustive set [S] of occurrent items; the so-called paradigmatically physico-dispositional, or "PPD" items. Taking the reductive thesis to be that there are no qualia in addition to those items, we can then see whether any of the dualist's strategies suggests otherwise. Thus, we shall take it that there is a *prima facie* problem for reductive physicalism if it can be shown that there are qualia not included in [S]. If there are, we shall then have to find out whether the qualia thus brought to light can be accommodated nevertheless by our commonsense account of the physical.

In view of the rather complex nature of the ensuing discussion, then, it might be as well at this point to summarise the main purpose and findings of the various chapters.

Chapters I - III.

We consider at the outset one of the most obvious ways in which the reductive physicalist might forestall any qualia-based challenge to his thesis; by claiming quite simply that there *are* no qualia to be accommodated. In chapter I we explore some of the ways in which it at least seems plausible that we might be mistaken about our sensory experiences, and clarify the eliminativist's apparently radical thesis that we are mistaken even to suppose that they occur. For clarity, the thesis is expounded initially in terms of particular types of *physical* object. We explain what it would amount to to take an eliminative stance with regard to those objects. Once the general concept of eliminativism has been thus clarified, we move on in chapter II to see how it might be applied to experience and qualia in particular.

In the second chapter, then, we explore the credentials of the eliminativist's suggestion that there are no such experiential qualities, or indeed, no conscious experience, to be explained. While the question of their occurrence remains open, we refer to

them as "intentionally inexistent" phenomena; phenomena to which we might attach intelligible predicates without thereby implying that instances of such phenomena actually occur. We might be able to understand what it would amount to for something to be a unicorn, for example, irrespective of whether or not there are any such creatures. Similarly, we might suppose that we can understand what it would amount to for a property to be a quale, or for someone to experience a quale, irrespective of whether or not we experience any such properties. As we shall see later on, what it *must* amount to to be a quale will depend on the dualistic strategy adopted.

Dennett, for example, tries to show that the positing of conscious experience and experiential qualities poses questions for which, even in principle, there is no determinate answer, and therefore that there is no justification for positing their occurrence. We argue that the same objection can be levelled on parallel grounds against his proposed alternative account in terms of computational or functional states or dispositions. We argue that Dennett's proposed strategy of simply treating what we remember having been experienced as fact [his so-called "operationalism"] is equally compatible with the positing of experience and qualia. Since, for the physicalist, the positing of such phenomena in order to account for the occurrent physico-dispositional traits is merely *redundant*, such facts are not logically incompatible with the occurrence of those phenomena. The sort of evidence which *would* justify the positing of conscious experience and qualia is logically independent of the physico-dispositional evidence. The strongest claim Dennett is entitled to, then, is that the positing of such phenomena is merely *redundant* as a proposed explanation of the acknowledged physico-dispositional [PPD] facts about sensory experience. But this leaves open the possibility that their occurrence might be established on some other, independent grounds, and even that they will satisfy the criteria for annexation on to [S]. As we shall observe later, even Dennett's admission that experience and qualia at least *seem* to occur is sufficient to create conceptual problems for his eliminativist stance.

The aim in chapter III is to explore the supposed distinction between eliminative and reductive physicalism in a more depth. In an attempt to understand the difference between the two theses we acknowledge at the outset that the predicates "*occur as physical*

properties discernible in introspection" and "do not occur" are at least intelligibly distinct. The problem then is to determine *which* properties are being attached to these predicates by each theorist. For unless we can establish that each is referring to one and the same property, their resultant positions are not even intelligibly distinct. The issue of whether an item occurs or not can only be a real issue if the item in question can at least be specified intelligibly. The aim, then, is to find some characteristic or property X which each would agree is the diagnostic property of qualia. Their respective theses will then be intelligibly distinct. QR will be claiming that properties which have property X; properties of type-[X], are PPD properties discernible in introspection, while QE will claim that properties of type-[X] do not even occur. But this is seen to be nothing more than a dispute over the appropriate *description* of qualia; in the absence of any further information about qualia, it seems acceptable to characterise the dispute as a disagreement over the appropriate description of the types of phenomenon which *are* discernible in introspection. And this leads us quite naturally into the ensuing chapters. What sort of description would the dualist *need* to produce in order to present a problem for the reductive physicalist?

It is acknowledged by at least some advocates of qualia-eliminativism [e.g., Dennett, on occasions, and Rorty] that there do at least *seem* to be qualia and experience, but that this amounts merely to our believing or judging wrongly that this is the case. The difference between QE and QR is then that only QR subscribes to that belief. What we find, however, is that casting seeming phenomenology in terms of belief *per se* affords no progress. For if the difference between QE and QR is just that while QR believes that he experiences qualia QE does not, there is still no intelligible account available of the *content* of the belief; the intentionally inexistent experience and qualia to which QR might refer. For further enlightenment, we turn instead to the claims made by the qualia-*dualist*. For even though no sense has yet been made of the dispute over whether qualia occur, we might gain further insight if the claim that qualia occur as non-physical properties of experience can be rendered intelligible. The remainder of the thesis explores various attempts to explain and justify the dualist's resultant position.

Chapter IV. The Inverted Spectrum Argument.

This chapter sets out to corroborate the dualist's claim that there do, indeed, at least *seem* to be experiential qualities[5] of a non-physico-dispositional nature and therefore, by implication, that the [intentionally inexistent] qualia we seem to experience cannot be characterised in physico-dispositional terms. The claim that qualia are *conceptually* distinct from dispositional states is supported by the standard version of the inverted spectrum thought experiment. The claim is that it is possible to imagine coherently that the set of *reactive dispositions* in terms of which an experiential belief is to be defined can vary in the presence of the given belief. My own conclusion is that this is an unjustified position, even if the conceptual import of expressions like "experiential belief" and "dispositional complex" is construed in a narrow sense. It seems intuitively obvious that there is an intelligible sense in which the content of our experiential beliefs can vary against a fixed set of simple *behavioural* traits; but it does not follow from this that the same can be said for dispositions. Thus, if a disposition is characterised as the behaviour which *would* be exhibited in some standard conditions, we can perhaps still maintain plausibly that spectral inversion with respect to dispositions of this sort might not be possible. Thus, although any dispositional account which presupposes an understanding of "standard conditions" might be faulted on independent grounds, neither qualia nor experience per se can be shown by the inverted spectrum argument not to be dispositionally definable. If, for example, standard conditions include Smith *wanting* to achieve certain ends, and *wanting* proves to be incapable of physico-dispositional analysis, the analysis will be false anyway and the inverted spectrum argument will be redundant. A parallel problem exists for experience. For when we believe that we have an experience containing Rp, there is no obvious way of showing

---

5. We refer here to qualia as "experiential qualities" without implying that qualia are to be taken literally as properties *possessed* by an experience per se. We have, as yet, no conceptual apparatus in terms of which to justify this claim. The more cautious claim is rather that there is experience, and qualia feature at least in part as the contents of experience.

that the belief itself does not co-obtain invariantly with a complex physico-dispositional phenomenon. Since the inverted spectrum argument fails to present even a prima facie problem for the reductive physicalist, then, there is no sense in trying to work out how the latter might try to accommodate the difficulties presented.

One further point of interest is that even if a prima facie problem were in evidence, our commonsense objective version of physicalism could survive. For even if the required spectral inversion were in evidence, it still would not follow that the contents of our qualia beliefs are not objective properties *per se*. For the inverted spectrum possibility against a fixed *physico*-dispositional backdrop would then be rendered unintelligible. It cannot be possible for it to occur if qualia are themselves objective, and hence physical, properties.

Chapter V. The Knowledge Argument .

The traditional version of the argument can be construed as a further attempt to demonstrate that qualia are occurrent, yet non-physico-dispositional in nature. The basic claim here is that since it would be possible to know all the physico-dispositional facts about seeing red, for example, and yet not know what it is like, qualitatively, to see red, the latter must be a nonphysical fact.

My conclusion is that taken in isolation the argument is unpersuasive. An opponent of reductive physicalism might insist that there is no a priori reason to suppose that all physical characteristics must be fully *teachable* by interpersonal demonstration and explanation of the sort permitted by the argument. On the other hand, if there were such a reason which depended on some form of conceptual private/public distinction between mental and physical phenomena [and our original brand of reductive physicalism does, in fact, assume such a distinction], the knowledge argument would become redundant. Qualia would be deemed mental just if they were epistemically private, but the assumption that they *are* epistemically private would require independent justification. The physicalist has at least three possible ways of attempting to resist this claim.

Firstly, he might adopt the reductive/eliminative stance that there just are no epistemically private phenomena to be accounted for. What happens to someone when they see colours for the first time, for example, is not that they learn a new experiential quality but that they simply acquire a new epistemic state with regard to already acknowledged physical phenomena. The ability hypothesis argues essentially along these lines.

As a supplement to the first option he might argue that although we might pick out qualia introspectively without *knowing* that they are paradigmatically physico-dispositional properties, they are indeed such properties. To suppose otherwise without further argument would be to presuppose that what are picked out [topic-neutrally] in introspection are not in fact just physico-dispositional phenomena already acknowledged by science. This position is compatible with the first insofar as it is eliminativist with regard to any irreducibly non-physico-dispositional phenomena, and reductivist with regard to those actually detected in introspection.

Finally, he could argue that qualia are distinct from any physico-dispositional characteristics already acknowledged as occurring, but that they are nevertheless additional characteristics of a physico-dispositional nature. What, for example, would we make of an [imaginary] instrument which is capable of providing absolutely reliable information about another observer's conscious experiences? It seems that there is no prima facie justification for ordaining that anything non-physical would be revealed through such an instrument. At the same time, however, we have no prima facie reason for rejecting the possibility of such an instrument. Hence, we appear to have no reason for rejecting the possibility that qualia are additional physico-dispositional phenomena which *can* be taught intersubjectively.

In the final analysis, then, the knowledge argument can only succeed on the presupposition that (i) all physical properties are inter-subjectively teachable to a blind person, for example, and that (ii) there are occurrent phenomenal properties which are not so teachable. Assumption (i) might be unjustifiable or inappropriate, and (ii) simply begs the question as to whether phenomenal properties occur and are neither physico-dispositional properties which have already been acknowledged nor additional phenomena to be

incorporated [non-reductively] into that category. The question remains, then, as to whether there *are* any such occurrent features of experience. It seems that the knowledge argument must at least be supplemented with independent facts about qualia, but then there is a danger that if those supplementary facts were sufficient to establish the dualist's case the knowledge argument itself would be rendered redundant.

Further efforts are then made to establish a case for the occurrence of non-physical phenomenal properties by exploring the implications of certain *modal* considerations.

Chapter VI. Kripke's Modal Argument.

At the heart of Kripke's challenge to the identity thesis is the Cartesian intuition that when we experience a pain sensation, for example, *this particular* pain [instance of the property Pain; see footnote 3] is only contingently related to any particular physical [neuro-physiological] state or phenomenon. We might extend the intuition to apply similarly to neurally grounded, but dispositionally characterised, phenomena. This particular pain seems to be only contingently related, if at all, to any such phenomenon as characterised in dispositional terms. More generally, this pain seems to be only contingently physico-dispositional at all. The difficulty posed by Kripke is essentially that all identities are, if true, necessarily so in a metaphysical sense [it is impossible that a phenomenon should be other than itself]. How, then, are we to explain why we have the strong intuition that the identities in question are at best only contingent?

The core of my interpretation of Kripke's intuition is that, notwithstanding his claim to the contrary, it is in fact epistemically based. I argue that we are only able to observe that this pain might not have been an episode of C-fibre stimulation because we have yet to determine epistemically that it is, or because we can imagine it not having turned out to be so. To suppose otherwise is to presuppose that even if we had already established that the identity obtains the intuition that it is contingent would survive. But I can find no justification for this presupposition. For if the two are in fact identical, our present inability to see

that this is the case, or ability to imagine it having turned out not to be the case, must be attributed to epistemic factors. As we saw in chapter III, it is possible to explain how we can know that this is a pain sensation without also knowing that it is an episode of C-fibre stimulation by appealing to the thesis that physico-dispositional states or episodes are identified only *topic-neutrally* in introspection. To suppose that this explanation does not work for sensations of pain is to presuppose that introspecting a pain sensation involves more than just being in such a topic-neutral epistemic situation with respect to any physical states. But this implies that pain sensations are already known to be distinct from all physical states; the very fact we have yet to establish. And although our analysis leaves an explanatory gap [Levine, 1983] which we cannot readily envisage being able to fill, it nevertheless remains possible that it is correct.

Once the crucial intuition of contingency has been redrafted in epistemic terms, Kripke's challenge can be reconstrued as the question of how it might occur in the case of introspected experiential phenomena. Following a by now well-trodden path [e.g., Nagel, Hill, McGinn] I concede that the epistemic asymmetry between introspection and paradigmatically scientific modes of observation *can* provide the required explanation, *unless* it can be shown on independent grounds that something more is involved in introspection. But if something more is involved, the "pain quale", or "what it is like" to introspect pain, for example, then that something is already irreducible to the physical phenomenon with which it was supposed to be identical. There is no further case to answer and Kripke's argument becomes redundant. If there is no such pain quale, however, Kripke's intuition does not even pose a threat to reductive physicalism.

There is, however, one possible rejoinder to this line of reasoning. If qualia do actually occur in introspection, it might be argued, they might nevertheless be identical with physical phenomena because the latter are identified in scientific procedure only topic-neutrally. Russell, Lockwood and Foster are notable proponents of this hypothesis. So even if qualia do occur as identifiable phenomena in introspection, they might turn out on a posteriori investigation to be the very causes of the physical effects via which we pick out C-fibre stimulation, say, in paradigmatically

scientific style. It is beyond the scope of the present thesis to explore this suggestion in depth, but we can at least cite some difficulties which it seems likely to encounter. The important point in the present context, however, is that even if it turns out to be implausible essentially no progress has yet been made in the attempt to discredit physicalism. There can only even be a prima facie difficulty for physicalism if there is some reason to suppose that our epistemic explanation for Kripke's intuition leaves something out, but this has yet to be established. Far from providing that reason, the position occupied by Russell et al. appears to presuppose that there are introspectible phenomena which have certain characteristics which paradigmatically physico-dispositional phenomena are not *known* to have. Unless this latter claim can be supported independently, then, there is nothing here for physicalism to explain.

Even if we concede that there obviously is something more to sensory discrimination than the mere topic-neutral discrimination of already acknowledged physical states, however, that "something" might still be an *objective* feature *per se*. And as in the previous chapters, we can see that Kripke's argument has nothing to say about this possibility.

Chapter VII. The Property Dualism Argument .

White confronts the epistemic version of the Cartesian intuition head-on. He argues that since, for him at least, any form of dualism is unacceptable, the only available options are analytical and "a posteriori" functionalism. Thus, while he concedes that what is picked out in introspection by a Smith who has a headache, for example, might turn out to be just a neural phenomenon, since this can only be known as such a posteriori, the property through which it is picked out epistemically in introspection must itself be a non-physical property. Since any form of a posteriori mental/physical identity thesis appears to entail the occurrence of non-physical properties, then, White offers what he considers to be the only available way out for the reductive physicalist. What Smith picks out in introspection, he argues, must be a *dispositionally* characterised, but neurally grounded, state or episode. His claim is that we can know a priori that our headache is such a dispositional

or functional state even though we can only discover a posteriori that it is grounded in some neural state or other.

In order for this part of his argument to succeed, however, our example of a physico-dispositionally humanoid robot establishes that further information about the phenomena actually discerned in *human* introspection is required, and therefore even White's argument essentially begs the question as to whether there is any such information available. My response to the property dualism argument is therefore that White's argument fails even to cite a prima facie problem for the reductive physicalist's position. As in the case of Kripke's argument, it is either ineffectual, because no problematic introspectible properties have been identified, or redundant, because White's alternative solution of identifying such properties conceptually with *dispositionally* characterised physical phenomena is clearly unavailable. Thus, we are left with the original question of whether any phenomena occur which cannot be construed as topic-neutrally discerned physical states or properties. If there are any such properties, it appears that they will resist all attempts to sustain any version of the reductive identity thesis other than the Russellian topic-neutral approach to physical phenomena, but we have yet to establish that they occur.

In the conclusion, we review the inadequacies inherent in the various attempts to refute reductive physicalism and, in the light of our overall findings, attempt to draw a clearer picture of what they would need in order to succeed. In particular, we look more closely at the physicalist's "topic-neutral" account of the knowledge we acquire in the first person, and consider how it might be shown to be false.

Chapter One

PHENOMENAL PROPERTIES - REAL OR ILLUSORY?

Any enquiry into the plausibility of the reductive physicalist's thesis, that the so-called mental items [objects, properties, events, etc.] are identical with purely physical items, must begin by considering whether there *are* any such mental items. One commonly employed approach looks at the elements involved in the conscious appreciation of secondary qualities. When we see something red, it is argued, we become aware or conscious that the experience has a distinctive quality which is quite different from the corresponding experiential quality associated with seeing something green, for example. The experiential quality which we ordinarily associate with the experience of something looking red, and which it is tempting to think enables us to determine whether it looks red, is then cited as the phenomenal property or "quale" which the physicalist is obliged to recognise and incorporate into his account of the world. At this stage it must be pointed out that the word "quale" is not being used in any particular technical sense, but merely to refer, even if, rather vaguely, to the experiential quality associated with seeing red , or what is often described [e.g., Jackson, 1986, p 291] as "*what it is like* to see something red".

One problem with this approach has been that such statements as that something *looks red* can be construed in a number of distinct ways. A parallel ambiguity occurs over the interpretation of Jackson's "what it is like" to see something red. In particular, there has been a tendency in the literature to equivocate between what Shoemaker [1981] characterises as the "intentional", or "relational", and "qualitative" interpretations respectively. The relevant distinction is picked up on and clarified by Ned Block when he draws attention to what he calls "the fallacy of intentionalising qualia"; the fallacy of assuming that the experiential quality associated with seeing red can be given a complete account in terms of *which objects* look red[6], or of which physical colour property

---

6. Such objects need not exist. Thus, a red unicorn is an

presents to the subject as looking red. Thus, according to Block [1990, p 54], the two interpretations of "looks red" and, by extension, also of "what it is like to see red", are:

1. The intentional interpretation.

   This involves the way experience represents or relates to the world. Since for each of us blood looks [with respect to colour] like standard red objects, then in the intentional sense blood looks red for both of us [i.e., with respect to colour looks like the same standard objects for both of us]. By the same token, "what it is like to see something red", for example, can be interpreted as being the same for each of us. What it is like to see red is interpreted just in terms of *which objects it is like seeing* [with respect to physical colour] to see something red [e.g., it is like seeing blood, ripe tomatoes, etc.].

2. The qualitative interpretation.

   This involves the *experiential quality* of what it is like for something to look red, or of seeing something red. If what it is like for you to see a standard red object is what it is like for me to see a standard red object, then looking red is qualitatively the same for both of us.

This distinction should be clear for our present purposes. If we assume that for something to look red to any particular observer is for it to produce a diagnostic experiential quality, or quale, the two interpretations of "looks red" will have the following implications respectively.

<u>The Intentional Interpretation</u>.

Suppose that for a particular object to look red to Jones is for that object to produce a particular and, for Jones, diagnostic quale Rp-[Jones] which standard red objects produce in standard conditions. It then follows that when presented with a new object it can be said to look red to Jones just if it produces Rp-[Jones].

---

intentionally inexistent object which looks red.

Similarly for other subjects. Thus, for Smith standard red objects look red just in virtue of producing a particular, and for Smith, diagnostic quale Rp-[Smith]. A new object can then be said to look red to Smith just if it produces Rp-[Smith].

For many commentators the temptation to pin down a particular quale in terms of intentional content has proven irresistible. Thus, according to this approach, and following on from what we have just supposed, an object is said to "look red" if it produces the quale produced by standard red objects, irrespective of the experiential character, or quale, associated with looking red for each observer. But this is to commit Block's intentional fallacy, since we have no guarantee that this experiential character is identical for each observer. According to Block, the fact that something looks red [intentionally] does not *entail* that it produces a particular quale in each observer.

The Qualitative Interpretation.

The problem is that if looking red for all observers were taken to amount just to producing some standard quale Rp, the possibility that Rp-[Jones] might differ from Rp-[Smith] would have been logically precluded. For if looking red *in general* amounts just to producing Rp it follows that Rp-[Jones] and Rp-[Smith] must both be identical with Rp. In order to preserve the logical possibility that standard red objects might look qualitatively different for Smith and Jones respectively, then, it follows that we are prohibited from defining any particular quale in purely intentional terms. In short, we are unable to define a quale Rp in general as that quale experienced when something looks like a standard red object, since looking like a standard red object might be qualitatively different for different observers. Thus, the setting sun might look red [i.e., look like standard red objects] to both Smith and Jones and yet the diagnostic quale experienced by each observer might be quite different. This is possible just because what it is like, qualitatively, for Smith to see standard red objects might not be what it is like for Jones to see standard red objects. This is what Block is getting at when he draws attention to the "fallacy of intentionalising qualia". To assume that it is possible to define or uniquely pick out a particular quale Rp as the quale associated with

looking red in general is to *presuppose* that looking red is not qualitatively different for each observer.

At this point it is important to note that any confusion between the intentional and qualitative properties associated with colour appearance is likely to render the reductive physicalist's position more plausible than it might otherwise be. It offers the possibility of concentrating on comparisons between objects, in respect of colour, at the expense of considering *what it is like* qualitatively to see particular colours. Colin McGinn, for example, is at pains to emphasise that once we accept the broadly Lockean dispositional analysis of colour, according to which for an object to be red is [roughly] for it to be disposed to produce certain sensory experiences in the observer,

> ...the essential point is that, according to the dispositional thesis, the ultimate criterion for whether an object has a certain colour ... is *how it looks* ... to perceivers. [McGinn, 1983, p 8; my emphasis].

The point is that despite this apparently unambiguous admonition, we might nevertheless fall into the trap of supposing here that McGinn means us to understand "how it looks" to perceivers in the intentional sense; as a reference to what class of objects in the physical world it looks like in respect of its colour. If the setting sun looks like ripe tomatoes, blood, a Santa Claus outfit, visible light of the longest wavelength, etc. [all viewed in standard conditions, of course], then it looks red just in virtue of looking the same colour as those standardly red objects. Resemblance in that sense, however, is entirely topic neutral with respect to the experiential quality characteristic of seeing red, or of something looking red. Indeed, it even leaves open the possibility that although we are capable of distinguishing between red and green objects, for example, there are no experiential qualities, or qualia, in virtue of which we are able to do so [one might imagine a rudimentary spectroscopic device, for example, which although capable of comparing and discriminating colours as effectively as we do, has no *experience* of qualia or anything else]. If the physicalist is allowed to take this line, then, life is made easier for him simply because he has escaped the need to acknowledge qualia and accommodate them within his account of the world.

On the assumption that looking red does amount to producing an experiential quality of some sort at least, then, [we refer to this here for convenience as a "quale"] the physicalist is obliged to accommodate that quale plausibly into his account. One way of drawing attention to this obligation is by envisaging the following possibility. It is [logically] possible to imagine waking up one morning to find that although all standardly red objects still look roughly alike in respect of colour, they all now appear qualitatively the way standardly green objects used to appear. Hence, the setting sun still looks red in the intentional sense, but the intrinsic experiential quality of seeing red in general has changed radically [this possibility is dealt with in detail in Chapter IV]. The point is that if we are to avoid falling into Block's "intentional fallacy" we must preserve the possibility of interpreting the dispositional thesis in this second way; how an object looks to perceivers might be construed as what intrinsic experiential quality it is disposed to produce in observers. According to this interpretation, if Smith finds one morning that standardly red objects look qualitatively the way standardly green objects used to look to him then the red objects simply *look green* to Smith on that occasion. So bearing this possibility in mind helps to remind us that there *are* [ex hypothesi] experiential qualities which the physicalist is obliged to accommodate within his account of the world.

Topic-Neutral Accounts of Colour

Failure to acknowledge these quite distinct interpretations of "how it looks", or "what it looks like", has already led a number of philosophers into proposing incomplete accounts of colour perception. J.J.C. Smart's claim that to have a yellowish-orange after-image is to have a visual experience as of a yellowish-orange patch, for example [Rosenthal, p169], might turn out to be substantially accurate insofar as the two experiences are *qualitatively alike*, yet incomplete because it says nothing about the intrinsic experiential quality itself. The danger here is that when we come to ask how the experiential quality might be accommodated within a physicalistic account of the world there is a temptation to conclude that there is no such quality to be accommodated. Smart himself seems quite content on occasions to

avoid the intrinsic content of experience altogether and explicate "looking green" in purely intentional terms.

> To say that something looks green to me is simply to say that my experience is like the experience I get when I see something that really is green. [Rosenthal, p 174] [7]

If we were to assume that there just are intrinsic experiential qualities, or qualia, and that no further demonstration is needed to substantiate this assumption, the idea that colour discrimination could in principle proceed in the absence of such qualities [as in the case of the spectroscopic device] would be simply redundant in the human case. For in that case the fact is that our own colour discrimination *would* be accompanied and facilitated by our being "directly conscious" of characteristic experiential qualities [Foster, 1991, pp 20-21].

In general, we can note that any account of human colour vision which provides a topic-neutral explanation of colour discrimination and recognition has nothing to say about these qualities. If they are real, therefore, any such account must be incomplete. The reason for this is that any such account can be interpreted entirely in intentional terms. The recognition of the redness of an object, for example, becomes the recognition that the colour appearance of the object is like the colour appearance of standard red objects, while

---

7. But see p 170 et seq. re remarks on Wittgenstein. Smart acknowledges that there are sensations, but claims that "sensations are nothing over and above brain processes". It is not entirely clear whether he is thereby acknowledging that there *is* a qualitative content to experience or sensations, and that this content is a brain process, or merely that there are only sensations insofar as "sensations" is taken to refer to brain processes. In his response to objection 3 [pp 172-3], he points out that it is possible to compare sensations without having to specify in what respect they are similar or dissimilar. But this leaves open the question of in what respect they *are* similar or dissimilar. If phenomenal properties are being compared, the topic-neutral approach *per se* fails to provide a physicalistic account of those properties, so that further explanation is required.

the discrimination between red and blue is construed as the discrimination between red objects and blue objects. The intrinsic experiential qualities cannot be given a purely intentional account just because, ex hypothesi, they have a non-intentional aspect. Hence, the fact that an account of colour discrimination and recognition succeeds in encompassing all the intentional facts about colour perception tells us nothing about whether there are such intrinsic experiential properties.

The Cartesian Intuition

The intuition that there are experiential qualities [phenomenal properties] associated with each colour and that it is these which enable us to recognise and discriminate between the experiences is compelling. Indeed, it seems so obvious to some philosophers that colour vision is characterised by such phenomenal properties that to suggest otherwise amounts to a flat contradiction of the facts. John Foster, for example, seems content to counter such a suggestion merely by asserting the contrary thesis that:

> I am now directly conscious of having a certain kind of visual experience - one as of sitting at my desk with a piece of paper in front of me. And while I can envisage ways in which this experience might turn out to mis-represent my physical environment (after all, it might turn out to be an illusion or a hallucination), I cannot envisage how it might turn out to be, *qua experience*, unreal. [Foster, 1991, pp 20-21]

Now, there are at least four quite distinct respects in which Foster's judgement here might be subject to error, and it is instructive to enumerate these possible errors before going any further. In terms of our own example of something looking red, the possible errors we might encounter, and at least some of which Foster appears to acknowledge in terms of his own example, might be characterised as follows.

When Smith judges that an object looks red to him, he might be mistaken in judging or inferring that:

1. The colour experience he is [in fact] having amounts to seeing a red object out in the world.

2. The colour experience he is [in fact] having is qualitatively as of seeing a red object [or as of an object looking red].

3. He is having any colour, or even visual, experience at all.

4. He is having any *experience* at all.

Clearly, Foster allows the possibility that he might be subject to an error of the first kind. It might turn out, he says, to be an illusion or hallucination. We might suggest further opportunities for error here. He might, for example, have been fitted, under a general anaesthetic, with a virtual-reality device which presents the experience to Foster as of him sitting at his desk, etc. But it is not clear which of the other three types of error he believes it possible that he might commit. He is "directly conscious of having a certain kind of visual experience" which, in the case of our example, would presumably be an experience qualitatively as of something looking red. But this ensures that, contrary to error 4, he is at least having an experience. Thus, it is at least inconceivable to Foster that he might be wrong about having any experience at all. We might suppose, furthermore, that he considers himself to be immune from either of the remaining two types of error. Contrary to errors 3 and 2, we might suppose, he is certain that the particular experience he is conscious of having is a *qualitative colour experience as of seeing red.* Thus, according to this interpretation, it would be impossible for Foster to imagine any way in which he might be mistaken either in his conviction that he is having an experience at all, or in his belief that the particular experience he is having is the *visual* experience qualitatively as of *something looking red.*

Now while we might object that Foster's inability to imagine how he might be wrong about either the reality or the particular qualitative character of his experience of phenomenal properties cannot amount per se to a proof that his judgement in those respects is infallible, and hence nor can it justify his certainty about the correctness of his judgements, it is nevertheless instructive to consider whether there are any respects in which it might be impossible for him to be wrong about these matters.

Experiential Illusion.

1. Let us concede at the outset that it is quite clearly possible to be subject to errors of type 1. Thus, as Foster suggests, it is difficult to imagine how the possibility of hallucination or illusion might be ruled out absolutely. Someone who is unfamiliar with the laws of optics, for example, might wrongly take the stick partially submerged in water to be bent, while as a matter of physical fact it is straight. To take an example involving colour perception, suppose that Smith has been fitted with red contact lenses while he sleeps. His initial reaction when he awakes might be to judge that the [white] walls of his bedroom have turned pink. It is clearly uncontroversial to concede that errors of this type are possible. It would be absurd to suggest either that the stick is *actually* bent, but only while submerged in water, or that Smith's walls are *actually* pink, but only when he is wearing the red lenses. Furthermore, the possibility of this sort of error is independent of any particular theory of colour perception. Even our rudimentary spectroscopic device might be expected to produce readings which are incorrect in this first sense if some sort of coloured filter is allowed to interfere with the incoming light.

Even more interestingly, examples of the first type of error can be envisaged in which the judgements are, so to speak, "topic neutral" with respect to the physical facts being judged. Thus, suppose that Smith is presented with two squares of card, A and B, such that the sides of B are one-percent longer than those of A. Suppose, further, that Smith is unable to discern the difference in size by visual inspection. Evidently, the difference between the size of the cards is below his discrimination threshold. Presented with A and B, then, he is quite likely to judge incorrectly [and topic neutrally, since the specific size is not stated] that the two cards are of the same size. So in such a case we can say that he has made incorrect comparative judgements which are, in a sense, topic neutral. In other words, since the comparison is explicitly in respect of *relative* size and does not contain or imply any specification of the absolute size of the cards, the error of judgement is topic neutral with respect to the absolute size of the cards. We can infer that since A and B are not identical he has made an error of judgement of type 1 with regard to relative size. Furthermore, given the task of estimating the *absolute* size of each card, it follows that he will

be prone to error in this respect also. For even if he assigns the correct absolute size to A, if he then judges A and B to be of equal size it follows that he has misjudged the absolute size of B.

2. Errors of the second kind might be envisaged in respect of colour if we assume that there are experiential qualities, or qualia, associated with colour perception, and that it is possible to misidentify those qualia. Thus we might have a concept of how the white walls look [the type of quale they produce] in "normal" conditions, and base our judgement as to what colour the walls are on how they look under those conditions. We would then find it hard to imagine being mistaken about the type of quale we are experiencing. In normal conditions we would be certain that they produce a white quale, while with red lenses installed we would be certain that they produce a pink quale.

If the illusion of the pink walls amounts to the walls actually looking pink in the sense that the quale produced by looking at them is the pink quale [i.e., the quale experienced by Foster when looking at a pink wall in standard conditions], then, is it possible that an observer might actually experience a pink quale and yet judge that it is, for example, a white quale? Suffering an illusion in respect of the walls' *apparent* colour would amount to having the walls be judged, or in some respect seem, to produce the white quale while in fact producing the pink quale. But it is not immediately apparent that there is any sense to be made of *seeming to appear* white, rather than simply *appearing* white. The intuition behind this point of view is presumably that a wall which seems to appear white just does appear white [i.e., produces the white quale] and therefore that there is no plausible sense in which one might be mistaken about this. The concept of illusion so far developed incorporates only the actual appearance of an object; the quale it actually produces in the observer. As such, it affords no clarification for the idea that something might seem or be judged to take on a particular appearance and yet not actually take it on.

Similarly, we can explore the notion of *seeming to appear* in a certain way, as distinct from actually appearing in that way, without the need to appeal to qualia at all. Thus, referring back to the example of the straight stick which is partially submerged in water, we have already established that it might be judged to be

bent even though it is straight. This provided an instance of the first type of error to which the observer might be prone. But is there any plausible sense in which he might be subject to the second type of error, an error about the appearance of the object? That is, can we make anything of the suggestion that it might seem, or be judged, to appear bent and yet actually appear to be straight? Clearly, such a distinction would presuppose an intelligible concept of "*appearance*" in this sort of case [just as in the case of colour-perception we required the concept of experiential qualities, the character of which might then be judged either correctly or incorrectly]. But what would it amount to to describe a stick as "appearing to be bent" and yet being *judged* to appear straight, in the circumstances described?

We might imagine an optically educated observer to judge, when presented with the partially submerged stick, that the stick looks "like a straight stick partially submerged in water" [as suggested by J.L. Austin, 1962, p 49], while a less sophisticated observer might simply judge that it looks like a bent stick. This sort of indeterminacy as to whether a stick looks like a bent stick or like a straight stick in water might lead to a corresponding indeterminacy in the sort of judgement an observer is likely to form about the physical nature of the stick, and therefore an indeterminacy over the disposition of observers to commit errors of judgement of type 1. Such indeterminacy will then inevitably infect any attempts to describe the appearance of the stick in topic-neutral terms; by referring to objects or physical circumstances in which a similar appearance [whatever that happens to be] would be produced. This is not to say that the indeterminacy is incapable of resolution, but any adequate solution is bound to involve quite complex specification of the objective circumstances which would produce an appearance similar to that being experienced. Consider the well-known optical illusion in which a road receding into the distance appears to have converging sides. When a well-informed observer judges not only that it is parallel sided, thus avoiding errors of type 1, but also that it appears to be parallel sided, is he judging [correctly] that it appears the way parallel-sided but receding roads normally appear, or is he judging [mistakenly] that it actually appears to be parallel-sided while in fact it appears to be converging? Clearly, we need an intelligible concept of "appearing" in order to decide between the two explanations. From

the foregoing discussion, however, we might at least infer that *if* there is such a phenomenon as a particular appearance in this sort of case [just as, in the case of colour perception, we might agree that there are particular experiential qualities, or qualia], then it at least seems to make sense to say that, as in the case of qualia, it is possible to be mistaken in our judgement about which particular appearance an object is producing. It at least makes sense to suggest that although the parallel sides appear to converge, they might be judged to appear parallel.

A more relevant example of this possibility of subjective error involves colour perception. Thus, if a red spot is set on a blue background the observer will tend to judge that it is reddish-orange. Objectively, there is no alteration to the colour of the spot here since the red and blue areas do not physically interact in any way [contrast this with the more complicated objective illusion in which a straight stick looks bent - presents the appearance as of a bent stick - in water, as a result of unusual physical circumstances which actually distort the light path from stick to observer]. Nevertheless, there is a sense in which it *seems* to the observer that the red spot is reddish-orange, since he judges it to be so. Assuming that the observer experiences colour qualia, then, the crucial question remains as to whether the quale he experiences in this case is as of a red or a reddish-orange object [in standard conditions]. If red, then the error of judgement might be regarded as the result of a *subjective* illusion, since it amounts to a false conviction about the subjective appearance presented by the red spot. He actually experiences a red quale but judges it to be reddish-orange. If, however, the red spot produces a reddish-orange quale in the observer, then the error of judgement might be referred to as purely an *objective* error of type 1. The objective colour of the object produces a misleading appearance [a quale which it would not produce in standard conditions], but the observer's judgement about the subjective character of that appearance is correct. He is simply wrong in judging that the physical spot is, in fact, reddish orange. Returning to the second type of error to which Foster might be subject, then, we can now see that whether or not errors of this type are possible depends on whether or not what we have referred to as a "subjective" illusion might occur. Is it possible, in other words, for an observer to be wrong about how an object appears; the subjective character of the experience it produces?

Again, as in the case of errors of type 1, it seems possible to imagine examples of errors of type 2 which are in some sense topic neutral with respect to the quality being judged. Thus suppose now that Smith has three cards, coloured in subtly different shades of red [he can ensure that this is objectively the case by mixing paints in the appropriate proportions for each card. The paint applied to A is just any commercially available red paint, while that applied to B has a little blue mixed in with it, and that applied to C has a little more blue mixed in with it]. He then finds that although he is unable to discern visually any difference between A and B in respect of colour, and similarly for B and C, he *is* nevertheless able to discern that C is bluer than A. Hence, he is able to infer logically that when he judges cards A and B, or B and C, to have the same objective colour he has committed an error of judgement of type 1 regarding the relative objective colours of the cards. But then it follows either that he has also made a comparative error about the *subjective* qualitative experience produced by each, or that there is no specific qualitative difference between the respective experiences. In other words, in order to avoid being committed to a subjective error of judgement in this case he must allow that when he looks just at A and B, or just at B and C, they *in fact* produce qualitatively the same experience in him. But this implies that the qualitative experience produced by at least one of the two cards in each case has changed. But if, as Michael Lockwood predicts [Lockwood 1989, p 164], Smith will fail to notice any such change, the conclusion must in any event be that a comparative [topic-neutral] judgement with respect to relative hue has been made incorrectly. In each case the error is topic neutral in the sense that it can be committed, detected and described without a single specific hue or quale ever having to be identified. As in the topic-neutral example of type 1 errors, however, this also implies that errors about the specific subjective character [quale] produced by each of the cards is possible. Thus, even if he assigns the correct specific quale to his experience of A, if he then judges A and B to produce the same quale it follows that he has misjudged the specific quale produced by B.

The above considerations should not be taken as a conclusive demonstration that errors of type 2 can be made, since in each example it remains possible, no matter how unlikely, that the subjective appearance of the observed objects will be correctly

judged. It is at least logically possible that the quale experienced when looking at B varies according to whether it is being viewed alongside A or C. If all three cards are viewed simultaneously, however, it becomes even more difficult to imagine how the observer might explain his failure to distinguish between the qualia produced by A and B, or by B and C, respectively, without conceding that an error of judgement has occurred. For if at the same time A and B are judged to produce the same quale and B and C are judged to produce the same quale, even though A and C do not, an error of judgement seems to be logically implicated. The only way of avoiding this conclusion is by maintaining that the three comparisons are conducted at slightly different times, and that the quale produced by at least one of the cards changes from one time to another. Even if we allow this possibility, however, we are now at least in a position to *understand* the sort of claims we would need to sustain in order to establish that such errors are possible.

3. If it at least makes sense to speak of a red object as only seeming [being judged] to appear [produce the experiential quality as of] reddish-orange, then, it is tempting to suppose that it also makes sense to speak of a red object as only seeming [being judged] to appear coloured [produce an experiential colour quale] at all. If looking reddish-orange entails producing the relevant experiential quality [quale] in an observer, then looking coloured more generally entails producing some colour quale or other. On that account, seeming to produce colour qualia amounts to being judged by an observer to produce colour qualia. More generally, the fact that it seems to be like anything at all to experience colours visually amounts just to the fact that the observer has a false belief, or makes an erroneous judgement, to that effect. As we shall see shortly, this is broadly the line taken by certain eliminativist philosophers. Richard Rorty, for example, explains that:

> ...the appearance-reality distinction is not based on a distinction between subjective representations [i.e., colour appearances] and objective states of affairs; it is merely a matter of getting something wrong, having a false belief [about the objective state of affairs]. [Rosenthal, p 270]

And if making errors of type 1 amounts just to having false beliefs about the objective facts [about the walls being white, or the stick

being straight, for example, or even as Foster suggests, that the walls and stick are mere hallucinations], we might expect to be able to apply this sort of analysis in a similar way to errors regarding the *appearance* of objects. Dennett, for example, agrees "wholeheartedly that there seem to be qualia" but goes on to insist that "this reasoning is confused, however" [Dennett, 1991, p 372]. Now however we try to refine our account of the appearance of objects, whether in terms of qualia, experiential qualities, mental representations, or whatever, the point is that there seems to be an opportunity here for someone of Dennett or Rorty's persuasion to object that for something in the world to seem to appear coloured amounts just to the observer having a *false belief* to the effect that the object appears coloured. Thus, errors of type 3 might turn out to be possible insofar as an observer might judge an object to look red, and yet be mistaken even in judging that it looks coloured. It is difficult to imagine a good example of an error of this sort, but it is at least prima facie a logical possibility. It seems to be at least a logical possibility, in other words, that one might actually have, say, a tactile experience and yet wrongly judge it to be a [visual] colour experience.

Dennett introduces an experiment in which "prosthetic" devices are used to provide sensory input and we might want to suggest that a type 3 error of judgement can be envisaged in the circumstances described.

> Prosthetic devices have been designed to provide "vision" to the blind, and some of them raise just the right issues. Almost twenty years ago, Paul Bach-y-Rita (1972) developed several devices that involved small, ultralow-resolution videocameras that could be mounted on eyeglass frames. the low resolution signal from these cameras, a 16-by-16 or 20-by-20 , array of black-and-white pixels, was spread over the back or belly of the subject in a grid of either electrical or mechanically vibrating tinglers called 'tactors'. After only a few hours of training, blind subjects wearing this device could learn to interpret the patterns of tingles on their skin, much as you can interpret letters traced on your skin by someone's finger. The resolution is low, but even so, subjects could learn to read signs, and identify objects and even people's faces. [1991, pp 339-40]

This experiment leads immediately to the following question: were these subjects experiencing conscious vision or just some prosthetic substitute? More specifically, let us assume that when using the prosthetic device they did in fact judge that they were having an experience as of, for example, seeing red. The crucial question is then: did their conscious sensory experience really have the quality as of seeing red, or did their "seeming phenomenology" amount merely to making a false judgement to that effect?

As Dennett observes, the result of this experiment was certainly the production of perceptual experience of some kind. The information supplied to the subject's back or belly by the tactor array led him to display spontaneous and appropriate responses to the events which created that information. After some training, for example, the subject took evasive action when the camera zoom facility was suddenly activated, as if he had become aware of the objects being viewed as having lurched suddenly towards his head [the location of the video camera]. But was this artificially induced perception really conscious *vision*? As Dennett himself wonders:

> Did it have the phenomenal qualities of vision, or just of tactile sensation? [p 340]

The actual results available from such experiments are inconclusive, to say the least, but Dennett is entitled to speculate. He does so, with some relish. The most plausible answer, he thinks, is that the subject will eventually report that:

> "....it's very much like seeing. I now effortlessly act in the world on the basis of information gleaned by my eyes from my surroundings. .... Without the slightest hesitation I react to the colors of things, to their shapes, and locations, and I've lost all sense of the effort expended to develop those talents and render them second nature". [p 343].

Now while it must be admitted that Dennett's subject has made impressive progress in learning to acquire most of the reactive dispositions usually associated with normal vision, his newly acquired skills might still nevertheless be deficient in one crucial respect. We do not yet know whether he is disposed to judge or believe that his conscious experience is *visual* or *tactile*. We can at least concede the possibility, however, that even though the

subject undergoes tactile experience, he might falsely judge his experience to be visual in character.

4. Finally, if judgements about our supposed experiences might be at least logically distinguished from the actual nature of those experiences in all the ways described above, there is yet a further possibility. The possibility is that even errors of type 4. can occur. When we judge that we are having any experience *at all* it might turn out that we are simply entertaining a false belief. It is this possibility to which Rorty alludes explicitly in the above passage. Evidently, despite the prima facie logical possibility that this might be correct, we have already seen that Foster "cannot envisage how it might turn out to be, qua experience, unreal" [Foster, 1991, p 20-21], thus apparently dismissing the possibility that what it *is* impossible to envisage might be merely that we do not have certain [erroneous] *beliefs* to the effect that there are real experiences. In a similar vein, Galen Strawson argues, or rather, assumes, that at least some of the crucial beliefs cannot be wrong in any important sense.

> The sense in which we cannot be wrong ... that if it seems to one that one is having an experience then one must indeed be having some experience or other. One can try the thought that the state of affairs of one's having rich and complicated mental experience might not really obtain but only seem to obtain. But it is self-refuting in Cartesian style, because for it to genuinely seem that such a state of affairs obtains is already for such a state of affairs to obtain. [Strawson, Galen. pp 99-100]

Here again, since the "only seeming to obtain" hypothesis would *not* be self-refuting if Dennett or Rorty's belief-based account of seeming to have an experience turned out to be defensible, Strawson's assumption that it *is* self-refuting amounts to begging the question against the eliminativist. Yet it is surprising how often contemporary thinkers simply disregard the latter and adopt a position at the outset which effectively presupposes that experience is real. Thus, Strawson explicitly misses the mark when he says:

> It seems that some philosophers want to say that sensations are really just judgements. Let them, so long as they grant that the .... ordinary view makes no error about the qualitative or experiential ... differences [between various experiences]. As it stands, their [the

> eliminativists'] view seems to be one of the most amazing manifestations of human irrationality on record. [Strawson pp 52-3].

According to the [amazingly irrational] alternative view, then, when Smith judges that the spot looks reddish-orange [produces the subjective appearance as of reddish-orange] he is mistaken because the spot does not look like anything [produce a subjective appearance] at all. Of course, it must surely be possible for Smith to have a true belief about *colour*. He might be correct in judging that it is red, for example, provided that he is not thereby invoking a notion of colour as a disposition to produce certain types of conscious experiential qualities. The conviction that such "qualia" are experienced is, according to the eliminativist, simply a false belief. Again, it cannot be the eliminativist's intention to reject the notion of colour altogether; what he actually rejects is just the notion [and belief] that objective colour is, or entails, the disposition of an object to produce experiential qualities, or phenomenal properties, of any sort in the observer.

While this is undeniably a surprising view, then, it would be question-begging to dismiss it out of hand on the ground that experiences just are the way they seem to be. For the whole point is supposed to be that seeming to be a certain way is not in itself some sort of introspectible appearance [in which case seeming would indeed amount to experience] but rather merely a judgement or belief to that effect; a judgement or belief which, for the eliminativist, must itself be entirely devoid of experiential content. It is not that we expect the eliminative position to turn out to be defensible in the final analysis; rather just that it deserves to be recognised for what it is and dealt with appropriately. It is an attempt to explain how our common-sense judgements about the reality of experience might turn out to be false. To cite our common-sense judgements as evidence to the contrary, therefore, is simply to miss the point and effectively beg the very question being discussed. Thus, when Strawson asks:

> What is it to suppose that one might be completely wrong [about the reality of experience]? It is to suppose that although it *seems* to one that there is experience - for this cannot be denied - there really isn't any experience.

> But this is an immediate reductio ad absurdum. For seeming is already experience. [Strawson, p 51].

he is explicitly begging the question as to the true nature of seeming. If Smith seeming to experience qualia, for example, really is nothing more than Smith believing or judging that he experiences qualia, then seeming is not "already experience" after all. In the ensuing discussion we shall be exploring the possibility in more depth of errors of type 4; errors in which a subject believes that he has conscious visual [e.g., colour] experiences even though he has no conscious experiences at all.

The Eliminativist's Account of Colour.

If the characteristic "appearance" of a red object for the eliminativist is just the having of a belief that an object has a certain property, however, it is by no means clear which property he is referring to. Dennett's apparently shameless offering is that "we detect the properties we detect" [1991 pp 382-3]. More constructively, we might suppose in the most general case that Rorty is referring to a 'belief that the object is coloured, or more specifically, that it is red. Admittedly this tells us nothing about the nature of colour as a property of objects in the world. Irrespective of what the colour of objects in the world is supposed to amount to, however, it is evidently disposed to produce certain effects in an observer, and we need to know something about the nature of those effects. Evidently, a red object is at least disposed to lead the observer to *believe* that it is red, but unless some further explanation of what the term "red" refers to here is forthcoming such a belief remains unexplained. Whatever the term "red" refers to, however, presumably the belief that an object is red is, at least, neurally realised in the observer [although it might still be *characterised* in terms of a dispositional complex of some sort]. When Smith reports that a ripe tomato is red, he is actually remarking on some property of the tomato which, under certain "standard" conditions, disposes it to induce in him either certain beliefs [and perhaps other dispositions] or the neural states which realise those beliefs and dispositions. We might then assume that for Rorty and Dennett, when Smith reports that a ripe tomato looks or appears red, he is reporting either that the tomato

produces the beliefs and dispositions or that it produces the neural states which a red object would normally be disposed to produce in him under those same "standard" conditions.

But this leaves the crucial question unanswered: when Smith reports that an object appears red by virtue of producing in him certain experiential qualities such as "colour qualia" or "what it is like to see red", for example, what is he actually reporting, according to the eliminativist? Jackson, Foster and others consider these experiential qualities per se to be immediately available for conscious introspection and their reality as experiential qualities to be beyond question. Robinson sympathises at least to the extent of supposing that:

> It must initially strike us as absurd to claim that we are wrong to believe that we are aware of a certain more or less determinate and recognisable [phenomenal?] feature when we suffer a pain or have a visual image [Robinson, 1982, p 81].

This, as we have already seen, however, is by no means the consensus view among philosophers. For the eliminativist either they are talking about nothing at all [i.e., nothing existent] or they are talking about something existent [e.g., reporting a dispositional state or the state of his visual cortex] but using the referring expression "qualia" or "what it is like to see red" in a misleading way. But in either case we would expect some explanation as to how the error can occur and in what sense it *is* an error.

Eliminative Physicalism.

As we have just seen, the question of whether we, as human beings, have qualitative experiences of a particular kind when for example, seeing red, is underpinned by a more fundamental issue. That is, do we have qualitative experiences *at all*, irrespective of the category to which they belong? For clearly, if we do not have such experiences at all, the question as to whether they can be regarded as being non-physical is entirely redundant. The intuition of the dualist is that certain characteristics of our visual experience must lie beyond the realm of the physical, but in order to vindicate

that intuition he must establish firstly that we do have visual experiences and, secondly, that they have those characteristics.

Conversely, then, we have seen that it is open to the physicalist to forestall any argument for dualism at the first hurdle. If he can show that the proposed candidates for non-physical status are not even characteristics of our visual experience, since we do not even *have* visual experience, it follows that there is no further case for him to answer. According to the eliminative approach we have been considering, the claim that the physicalistic account of the world leaves nothing out then remains intact by default. The belief or judgement that Smith has an experience qualitatively as of seeing red amounts to just that; a belief or judgement which itself has no experiential content. Such a belief is deemed false by the eliminativist in that the object of the belief, the experiential quality, simply does not exist. The problem with this approach, however, is that it is by no means obvious exactly *which* items or qualities are thereby deemed not to exist.

Even before we raise the question of whether physicalism is true, therefore, we need to begin by attempting to provide an intelligible account of "eliminativism" as a general concept. What does it amount to to deny the existence of, say, items of type-X, rather than claim that they are being misdescribed in some way? In order to clarify this distinction, let us begin by assuming that Jones is an eliminativist with respect to items of type-X just if he believes that items of type-X [which, ex hypothesi, include experiential qualities, or "qualia"] do not exist.

Suppose that Smith believes that when he refers to his experience of red he is at least referring to something [Rorty's subjective representations, Dennett's qualia, for example] of type-X, even though the question as to the physical or non-physical nature of items of type-X has yet to be raised. As we have already seen, Jones's response to this position would amount to pointing out that, although Smith indisputably does have this belief or judgement, it is false, since items of type-X [and therefore qualia] do not exist. Thus, suppose that Smith has adopted the convention of referring to the phenomenal property associated with his seeing red with the expression Rp. Similarly, when seeing blue, assume that he refers to the diagnostic phenomenal property as Bp, and that he regards each

of these to be an item of type-X [e.g., an experiential quality, or qualel. The initial charge would then be that his belief that he is experiencing "Rp" , or "Bp", is simply false, since there are no items of type-X. Assuming that both Smith and Jones understand what it would be for an item to be of type-X, then, the difference of opinion seems well-defined. Smith believes that items of type-X exist and that he experiences them, while Jones believes that items of type-X do not exist and therefore that he does not experience them.

Unfortunately, however, the apparent crispness of this account is illusory. In particular, we have yet to discover whether Jones believes that Smith's experience of red is an experience of an item of some *other* type, [X'], or of *no item at all*. Let us consider these options in turn. When Smith expresses the belief that he is experiencing items of type-X Jones might take either one of the following to be the case.

> 1. Items of type-X do not exist, and Smith's reference to items of type-X is a reference to no existent items at all.

This position seems to be unequivocally eliminativist with respect to items of type-X in the sense that when Smith claims to be experiencing items of type-X he is indeed being held to be reporting the experience of items of type-X but such items are being held not to exist. His reference to items of type-X is simply a reference to no existent property or characteristic of experience.

> 2. Items of type-X do not exist, but Smith's reference to them is in fact a reference to experiences of items of type-X'.

Clearly, Jones is still eliminativist with respect to items of type-X, since he believes that there are no such items. But at the same time he believes that whenever Smith reports an experience of an item of type-X he in fact experiences an item of type-X'. In a sense, then, Jones might reasonably assume that Smith's reference to "items of type-X" must be an [inaccurate] reference to items of type-X'. The items he refers to exist, but are not exactly as Smith describes them [they have property X' rather than property X]. So in that case Jones is *not* an eliminativist with respect to the items referred to by Smith as "items of type-X". In case 2, therefore, we

need to clarify the concept of eliminativism with regard to the items postulated by Smith by making the following distinction. We need to distinguish between:

(i) Items of type-X,

and

(ii) Items *referred to by Smith* as "items of type-X".

As we have seen, it is perfectly intelligible to suppose that while Jones is eliminativist with respect to (i) he is *not* eliminativist with respect to (ii). He might believe that there are no items of type-X to be experienced, but that when Smith uses the expression "items of type-X" he is in fact referring, albeit misleadingly, to items of type-X', which, we might suppose, he does experience. Hence, the evaluation of any eliminativist position is only possible if it has been made clear with respect to which items the position is eliminativist. For example, while Jones denies that there are any qualia to be experienced, he might nevertheless accept that there are *neural states* which Smith does experience when he sees red [see Paul Churchland, 1989, chapter 3, for an example of this sort of position]. There might then be some uncertainty as to whether Smith's report of experiencing qualia should be interpreted as a report about experiencing neural states, rather than a report about nothing at all.

Fortunately, the brand of eliminativism we are currently considering seems to offer a way out of this problem. For not only does it claim that there are no *qualia* to be experienced, but that there is no *experience* at all. And if there is, ex hypothesi, no experience at all, then it follows that option 1 must be the correct one. For if there is no experience at all, it is not possible that Smith's expression "experiential qualities", or "qualia" might actually be referring to some other experiential item, since there are none. So provided just that the expression "experience" is understood, there is no problem about interpreting Jones's position; it is that he is eliminativist with respect to experiential qualities because he is eliminativist with respect to experience.

Suppose now, however, that we are *not* yet sure that we understand the expression "experience". In particular, we have yet to discover whether Jones believes that what Smith refers to as an "experience" is some other sort of physical phenomenon [a neural state, for example] or no occurrent phenomenon at all. Let us consider these options in turn. When Smith claims that seeing red involves having an "experience", Jones might take either one of the following to be the case.

> 1. Experiences do not exist, and when Smith reports having an experience he is having [undergoing] nothing at all.

But this is clearly absurd, since when Smith is seeing red, for example, even the eliminativist wants to accept that he is at least undergoing a neural episode of some sort or other. So option 2 is the only one available. thus:

> 2. Experiences do not exist, but when Smith reports an experience he is in, or undergoing, some neural state or other.

But once this is acknowledged, we again encounter the problem of how to interpret Jones's brand of eliminativism, this time with respect to *experiences*. Thus, we need to distinguish between:

(i) Experiences,

and

(ii) Items *referred to by Smith* as "experiences".

Once again, it is perfectly intelligible on assumption 2 to suppose that while Jones is eliminativist with respect to (i) he is not eliminativist with respect to (ii). He believes that seeing red involves no experience, but that it is at least accompanied by being in some neural state or other. Hence it is possible that when Smith refers to an "experience", Jones takes him to be in fact referring, albeit misleadingly, to a neural state [which, we might suppose, constitutes seeing red, for example]. Hence, an unambiguous interpretation of the eliminativist's position is again only possible if it has been made clear with respect to which items the position is eliminativist. In the present case, does he take Smith's

expression "experience" to be a reference to nothing at all, or rather misleadingly to some neural state or other which does exist?

Eliminativism and Referential Indeterminacy.

The problem is a familiar one. An eliminativist statement of the form "there are no items of type-X" [e.g., "there are no experiences"] can be interpreted in two distinct ways, either to the effect that whenever we use the expression "items of type-X" we refer to nothing at all [because there are no items of type-X] or to the effect that in such cases we refer to something, but are mistaken in implying that the referents *are* of type-X]. Thus, assuming that we subscribe to the statement that "Yetis do not exist" we might nevertheless be interpreted as believing either that "The creatures you refer to as bearing Yeti [Yeti-type] characteristics simply do not exist" or that "the creatures you refer to as bearing Yeti [Yeti-type] characteristics exist, but they do not in fact bear Yeti [Yeti-type] characteristics; they are only bears". The difference between the two interpretations in the present case appears, at least *prima facie*, to exemplify the difference between *eliminative* and[1] *reductive* physicalism with respect to the items we refer to as "Yetis". The impression given is that we are either eliminating Yetis by assuming that our expression "Yetis" refers to Yetis and then simply saying that Yetis do not exist, or reducing them to bears by saying that they [the creatures we refer to with the expression "Yetis"] are actually bears.[8]

---

8. There are complications, however. For example, assuming that we already know that there are bears in the region, the claim that "Yetis" in fact refers to bears implies that there are fewer species than we thought, and in that sense a species [i.e., the Yeti] has been eliminated from our ontology. So even whilst reducing, or translating, Smith's discourse about Yetis into discourse about bears, we are eliminating Yetis from our ontology as a separate species. This point will emerge as significant in the latter part of the present chapter.

More fully, assume that the thesis being contested by the eliminativist is that items of type-X are non-physical in virtue of bearing property Y. Thus, for example, items of type-X might be *experiences*, and Y might be the property of *not being encapsulated by physics*. There are four distinct responses the physicalist might make:

1a. Items of type-X exist, but do not bear property Y.

1b. Items of type-X exist, and bear property Y.

or:

2a. There are no items of type-X, but you are using the expression "items of type-X" misleadingly to refer to items of some other type.

2b. There are no items of type-X, and you are using the expression "items of type-X" to refer to nothing at all.

Substituting "Experiences" in for "Items of type-X", and "are encapsulated by physics" in for "bear property Y", the above options become:

1a. Experiences exist, but are encapsulated by physics.

1b. Experiences exist, and are not encapsulated by physics.

or:

2a. Experiences do not exist, but you are using the expression "experiences" misleadingly to refer to items of some other type, [e.g. neural states] which *are* encapsulated by physics.

2b. Experiences do not exist, and you are using the expression "experiences" to refer to nothing at all.

From this we can see that the various options amount respectively to the claims that:

1a. Experiences exist but they are physical in nature.

1b. Experiences exist and they are non-physical in nature.

2a. There are no experiences, but you misleadingly use the expression "experiences" to refer to, for example neural states, which *do* exist.

2b. There are no experiences, and you use the expression "experiences" to refer to nothing at all.

For the purpose of the present discussion options 1a and 1b are relatively unproblematic. Option 1a claims that experiences are physical. We should note, however, that it says nothing about whether or not the resultant physicalistic position is reductive or non-reductive [i.e., whether or not experiences are reducible to paradigmatically physical constituents]. For the sake of clarity, therefore we might adopt the term "incorporative" to indicate that irrespective of which version is intended, experiences are claimed to be incorporated into physics. Option 1b seems to be an uncontroversially anti-physicalist position about which we need say nothing more at this point. The real difficulties arise over options 2a and 2b.

The problem, in short, is this. If the difference between 2a and 2b is essentially the difference between taking the referent of Smith's term "experience" to be a neural state, for example, and taking it to be nothing at all, how are we to decide which of these readings is correct? In other words, is there any independent criterion for deciding what the intended referent of an expression happens to be for a particular speaker?[9]

According to Richard Rorty, the appropriate response to this sort of question is fundamentally indeterminate and might be influenced, for example, by how accurately the existent items mentioned in option 2a are being described [Rosenthal, p272]. Thus, irrespective of whether it is qualia or experiences that are in question, we might choose the interpretation in 2a by reasoning that:

---

9. See Robinson, 1994, pp 2, 72, for an account of "intentionally inexistent objects".

2a. Incorporative Physicalist Response.

> Qualia/experiences, as you describe them, do not exist. You are talking about dispositional/neural properties or states, but some of your claims about them are false [i.e., the items you refer to as "qualia/experiences" and I refer to as "dispositional/neural properties or states" are one and the same, but you are describing them incorrectly as, for example, being non-physical, or epistemically private].

Essentially, then, this response would be equivalent to the claim that what we refer to as "qualia" *do* exist but that they are, for example, physically realised but dispositionally characterised states rather than conscious experiential qualities. Similarly, what we refer to as "experiences" do exist but they are, for example, neural states rather than conscious episodes. As such, claim 2a turns out to amount to much the same as claim 1a, with the additional condition being added about which word we should be using to refer to the existent referents [e.g., neural/dispositional states or belief states]. As we saw, this brand of physicalism can be either reductive [if the existent referents are constituted by paradigmatically physical items], or non-reductive [if they are not].

Alternatively, we might choose the interpretation in 2b on the grounds that:

2b. Eliminative Physicalist Response.

> Since practically nothing you say about "qualia/experiences" is true of dispositional/neural states, you must not be talking about dispositional/neural states [and your expression "qualia/experiences" has no existing referents in the world].

From this possibility it emerges that the eliminative approach, while amounting to the claim that certain items do not exist, is likely to run into difficulties when it comes to saying *which* items do not exist. In the [case] we are concerned with the physicalist might claim either that "qualia" exist but are physico-dispositional states, or that they have [intentionally inexistent] referents of which there are no instances in the world. All he is saying in

effect, then, is that *if* "qualia" is intended to refer only to non-physical qualities there are no qualia. But *all* physicalists will subscribe to this statement. If, on the other hand, he takes it to refer just to experiential qualities per se, then qualia might exist as physically constituted items. It all depends on what we mean by "qualia". Thus, one way of understanding the eliminativist's position is by noticing that *all* physicalists are eliminativists with respect to certain [intentionally inexistent] items but not to others. The fundamental problem of understanding an eliminativist statement consists just in identifying the items whose existence is being denied and, hence, the ontology of the physicalism subscribed to.

The Elimination of Qualia.

The main problem is that in general qualia are not assigned a sufficiently clear identifying property X and therefore it is seldom clear whether the physicalist's treatment of qualia amounts to an incorporation into or an elimination from the physical ontology. The physicalist is free to interpret his own position in either of these ways depending just on whether he is prepared to be charitable or not. He can either assume, charitably, that the qualia discourse is about items which do exist; dispositional or neural properties, say, but is wrongly being described as being about something else [as in "you are right to say that Santa Claus exists, but in fact he is your father"/ "you are right to say that the items you refer to as 'type-X items' exist but they are dispositional or neural properties"], or less charitably that it is about something else which does not exist ["you are wrong to say that Santa Claus exists; it was your father"/ "you are wrong to say that items you refer to as 'type-X items' exist; there are only dispositional or neural properties"]. If "qualia" is used in such a way that it is possible to be charitable [i.e., it is sufficiently consistent with the physical facts , as in "it is possible to interpret your expression 'Santa Claus' as referring to your father, since almost everything you say about him is true of your father"/"it is possible to interpret your expression 'items of type-X' as referring to e.g., dispositional or neural properties, since almost everything you say about them is true of dispositional or neural properties"], then the physicalist's response might reasonably be construed as *reductive*

[provided that the neural or dispositional properties concerned are already agreed members of [S]]. On the other hand, if the use of "qualia" cannot be interpreted in such a way [as in "it is impossible to interpret your expression 'Santa Claus' as referring to anyone who exists since so much of what you say about him (he lives at the North Pole and flies through the air on a sleigh pulled by reindeers, etc.,) is true of no-one"/"it is impossible to interpret your expression 'items of type-X' as referring to anything physical, since so much of what you say about them (they are experiential qualities, etc.), is true of nothing physical"] then his response is eliminative.

But, as we have seen, this leaves us with the problem of how wrong we can permit someone to be and yet still take him to be referring to a particular item. Smith might be construed as saying something *false* about an item which nevertheless exists [for example, " 'Santa Claus has many helpers' refers to Smith's father but is false"/"'qualia are epistemically private' refers to dispositional or neural states but is false"]. Clearly, on the other hand, the same assertion could be reconstrued as saying something *true* about an item which does not exist [as in "Santa Claus has many helpers" is true (in the myth), but he doesn't really exist"/"Qualia are experiential qualities" is a conceptual truth but they do not exist].

Thus, referring back to our discussion of Foster, the indeterminacy comes out in the following way. We saw that Smith's judgement that something "looks reddish-orange" might be construed as Smith judging that it produces the "reddish-orange quale", and that as such there are several respects in which he is at least logically vulnerable to error. In short, he might actually be experiencing the *red* quale and mistakenly judging it to be reddish-orange, or even be mistaken in his belief that he is undergoing any experience at all. And it is precisely in virtue of the logical distinction between Smith *experiencing* a particular quale and Smith *judging or believing* that he is experiencing that particular quale that there is room for error. For the physicalist, Smith might be construed as going astray in a number of distinct ways. The indeterminacy of reference described earlier renders it at least logically possible, despite his belief to the contrary, and despite assurances from such

philosophers as Robinson that the experiential quality of seeing red, or of having a pain, is:

> that aspect of the world which we have agreed the disappearance theorist [eliminativist] cannot be seriously intending to abolish [Robinson, 1982, p 84],

that when Smith refers to a specific quale he might be interpreted *either* as referring to a certain dispositional or neural state [charitable, reductive response] *or* as referring to nothing at all [uncharitable, eliminative response]. It is only the latter of these two interpretations, the one which Robinson finds incredible, which casts Smith as a proponent of items which do not exist, and Smith's interpreter as an eliminativist with respect to those items.

If a physicalist is to count himself as an eliminativist with respect to Smith's qualia, therefore, he is committed to the following. Firstly, he must claim that when Smith refers to a particular quale, Rp, for example, he is not referring to anything paradigmatically physical, such as a dispositional or neural property. Secondly, he must claim that there are no non-paradigmatically physical items, belief states for example, to which he might be referring and which might turn out on further investigation to be physically realised. Finally, and for whatever reason, he is not prepared to simply *add* Smith's qualia irreducibly to the items whose existence he already acknowledges. This leaves just the eliminativist's option, which is that Smith's "Rp" refers to nothing which Jones is prepared to incorporate into the physical domain and therefore to nothing which exists.

Even before getting into a debate as to whether qualia are physical items, then, the difference of opinion as to whether or not they exist can be described in the following way. Assume that there is some definable set [S] containing all and only the items acknowledged by the eliminativist, Jones, as being incorporated [reductively or non-reductively] within his account of the world [i.e., existing]. Smith might acknowledge the existence of all members of [S], and yet claim further that experiential qualities, or qualia, should be included in that set, either by reduction to an already acknowledged member of [S], or as an additional item. [10] In contrast, the eliminativist will claim that Smith's "qualia" are *not* reducible to any members of [S] and neither do they qualify for

incorporation as additional members of [S]. Then, since *ex hypothesi* [S] is exhaustive for Jones, it follows that for him the proposed "qualia" do not exist. Thus, the point of departure for the two views comes to light when we compare the two views of qualia in relation to [S]. Smith claims simply that the [intentionally inexistent] referent for "quale" exists. Jones claims that Smith's [intentionally inexistent] referent for "quale" is not to be incorporated into [S], either reductively or non-reductively, and therefore concludes that it does not exist.

This leaves only one additional condition to be attached to the eliminativist's position. It is that since he is a physicalist the set [S] contains only physical members. His requirement is that candidates for inclusion in [S] will only qualify if they are physical. If he deems Smith's "qualia" ineligible for membership of [S], then, he does so [presumably] because he does not, or cannot plausibly, interpret the items referred to as "qualia" as belonging to his physical account of the world. From Smith's point of view, of course, physicalism may or may not be true, so he remains free to choose either of the following options. He can either claim as a physicalist that his qualia should be incorporated into [S], either reductively or non-reductively, or as a non-physicalist that they should be acknowledged as existing even though they do not belong to [S].

This seems to be about as clear as we can be about the nature of eliminativism with regard to qualia, and yet there remains an apparently irresolvable indeterminacy in the account. For no matter how resolutely the eliminativist insists that Smith's "qualia" are not reducible to, or eligible for inclusion as, a member of his own set [S], the question remains as to how, or on what ultimate grounds, he is able to justify his verdict. As Quine asks:

---

10. We must keep these options open at this stage to avoid begging the question as to whether Smith believes that the reductive or non-reductive treatment of his "qualia" is appropriate; all Smith claims at this stage is that qualia *exist*].

> What now can we make of the difference between identifying the mental states with the states of nerves, as I just did, and repudiating them rather in favour of states of nerves? I see no difference. In either case the states of nerves are retained, mental states in any other sense are repudiated, and the mental terms are thereupon appropriated to states of nerves. So I may as well persist in calling my proposed reduction of mind to body an identification of mental states with bodily ones, neural ones; a construing of the mental as neural. [Quine 1985, Rosenthal pp 287-8].

In terms of what we have just been saying, then, Quine evidently agrees that the eliminativist's position is only intelligible insofar as the contents of his set [S] might be specified [e.g., states or properties of nerves qualify, but "qualia" not construed as states or properties of nerves do not]. But this is compatible with a standard *reductive* physicalism in which mental properties *are* taken to be neural properties. It leaves open the question of whether Smith's discourse about "qualia" *should* be construed as being about neural properties and thereby incorporated into the ontology. Hence, what makes a physicalist's position eliminativist with regard to qualia is just his decision to interpret reference to "qualia" as not being reference to any members of [S] [neural or dispositional properties, for example], and therefore, because ex hypothesi, [S] is complete, his decision to interpret reference to "qualia" as reference to nothing existent.

Eliminativism and Seeming..

Once the eliminativist's position has been formulated thus, as a decision to interpret discourse about qualia, or experience in general, as discourse about nothing which exists, a further complication can be introduced. If the eliminativist's claim is that Smith's qualia simply do not exist, the question then arises as to how it is possible for it to *seem to Smith* that they do exist. For if it really is the case that it seems to Smith that he is experiencing the reddish-orange quale, then according to the eliminativist's account of seeming it follows that he must at least have an *intelligible concept* of the reddish-orange quale, in virtue of which he is able to claim intelligibly that he seems to experience that quale. But it is not at all clear how the

eliminativist would make sense of the concept of some non-physical experiential property which, ex hypothesi, he has not experienced [because the property does not exist and there is no such phenomenon as experience].

Now, the important point is that viewed from Smith's point of view referential indeterminacy does not infect the *concept* of a specific quale as it infected *Jones's* interpretation of Smith's "qualia" and "experience". When he claims that something looks reddish-orange, and thus that he at least seems to be experiencing the reddish-orange quale, Smith might nevertheless still be quite clear about *what* he seems to be experiencing. Even though he might in fact be experiencing the red quale. Similarly, even if Jones is right and it turns out that Smith is experiencing nothing at all, Smith is at least able to claim intelligibly that he is experiencing something. A fortiori, then, if he finds these judgements intelligible he at least understands what it is like to have an experience as of the reddish-orange quale, and therefore has an intelligible concept of the reddish-orange quale itself as a particular quality of experience. In short, if Smith's claim that he seems to be having an experience of a particular type is intelligible, and Dennett suggests that it is, then the *concept* of an experience of that type is intelligible too.

But if it really is the case that Smith at least has an intelligible concept of experiencing the reddish-orange quale, in virtue of which he can claim intelligibly that he seems to experience that particular quale, then the eliminativist is obliged to accept that Smith is making an intelligible claim about a phenomenon which is non-occurrent. Now it is not at all clear how he would make sense of a concept of a particular non-physical experiential quality which, ex hypothesi, he has not experienced [because the property does not exist and there is no such phenomenon as experience]. And yet it is not only the opponents of eliminativism who claim to at least seem to be having experiences with particular qualitative characteristics; even some eliminativists evidently feel obliged to concede this much to common sense. Thus, Dennett, for example, explains that:

> There seem to be qualia, because it really does seem as if science has shown us that the colors can't be out there, and hence must be in here. Moreover, it seems that what is

> in here can't *just* be the judgements we make when things seem colored to us. This reasoning is confused, however. [Dennett, 1991, p 372].

What we must emphasise here is that the problem of how to explain how we can seem to experience a quality which does not exist falls squarely on the eliminativist alone. For even while experiencing qualia might turn out to be experiencing neural states [qua Quine, above], it remains a mystery as to how we can seem to be experiencing qualia and yet be experiencing nothing at all. But the difference between these two theses exemplifies precisely the difference between reductive and eliminative physicalism [or materialism] with respect to qualia. Construing talk about qualia as talk about neural states is explicitly reductive, while construing such talk as being about nothing at all is explicitly eliminativist.

Finally, suppose that the eliminativist claims that the observer does not even *seem* to experience a specific quale, the red quale, say, since the concept of that quale [or of any other, for that matter] is not even intelligible. There is, surely, at least a prima facie plausibility in the claim that discourse about qualia is simply unintelligible, and that it is in virtue of this fact that qualia can be said not to exist. The evaluation of this possible manoeuvre is the main theme of the next chapter. What we find, in short, is that qualia discourse cannot be shown to be unintelligible in virtue of *conflicting* with the known physico-dispositional facts about colour vision, and so must be so if at all just in virtue of being redundant; adding no descriptive or explanatory power to the physico-dispositional account. But if this is the case it follows that the eliminative option with regard to qualia must be unwarranted. If qualia discourse is unintelligible in virtue simply of being a redundant appendage to the physico-dispositional account such discourse does not conflict with that account in respect of any physico-dispositional facts. In the absence of any further information, therefore, it is permissible to construe qualia discourse as [somewhat misleading] discourse about the physico-dispositional facts. There is no reason why qualia discourse cannot be construed as just another way of talking about neural properties, for example. But this is the *reductive* option. Hence, it follows from the redundancy claim that there are no facts in virtue of which

qualia discourse is shown to be about nothing at all rather than about, say, neural properties.

Conclusion.

Being an eliminativist with respect to qualia or, more generally, to experience, essentially consists in taking discourse about "qualia" or "experience" as not being about occurrent physical properties or states, either neurally or dispositionally characterised and therefore, for the physicalist, not being about anything at all. In the ensuing discussion we shall be exploring, firstly, the possible reasons for adopting this position rather than the alternative and more charitable position of reductive physicalism.

Secondly, we shall be exploring the possibility of assimilating the eliminativist's position with the apparently incompatible concession that we do at least *seem* to experience qualia. On the one hand the eliminativist is convinced that the sum total of facts about the world must be facts about the items contained within the set [S] of all physical items and that qualia do not belong to that set. On the other, Smith insists that it is intuitively obvious that he at least has the intelligible concept of an item [the reddish-orange quale, for example] in virtue of which it seems to him that he experiences that quale, but which the eliminativist disqualifies from membership of [S]. The question we shall be exploring is whether the eliminativist's position on this point is justifiable, or even intelligible. Is it possible that it should really seem to someone that he experiences qualia, or have the intelligible belief that he experiences qualia, even though they do not exist? Even if it is possible, on what evidence is the eliminativist justified in construing Smith's beliefs about qualia or experience in general as beliefs about properties or phenomena which do not in fact occur?

Finally, we ask in the light of the above considerations whether it even makes sense for the eliminativist to claim that *he* does not experience qualia, or that he does not have any experiences at all.

Chapter II

ELIMINATIVISM AND REDUNDANCY

One conclusion of the first chapter was that Jones's elimination of qualia amounted to his decision to construe Smith's discourse about "qualia" as not being discourse about any items which Jones either already accepts as, or is willing to include as, members of his exhaustive set [S] of existents. And if he is construing Smith's discourse as being about no such items, then since for Jones membership of [S] is a logical prerequisite for existence, he must be construing it as being about nothing which he believes to exist. In terms of an *intentionally inexistent* referent, we saw that Smith's referring term "Yeti" might be construed either as referring to the Yeti, even though it does not exist and the term "Yeti" *might* not even be intelligible, or as referring, albeit misleadingly, to the bear, which does exist. Jones is an eliminativist with regard to Smith's "Yetis" and "qualia" just because he takes these expressions to refer to Yetis [rather than bears], and qualia [rather than, say, neurally realised (but perhaps dispositionally characterised) properties], and accepts the existence of neither. Our first consideration now is on what grounds this construal of Smith's discourse about qualia or even about experience might be *justified*. Secondly, if it should then turn out that their existence might justifiably be denied, can it be justifiable, or even make sense, to then concede, nevertheless, that they do at least *seem* to exist?

Dennett's Unverifiability Thesis.

Perhaps the simplest and most direct way of supporting the eliminativist's interpretation of discourse about phenomenal properties or qualia is by showing that their existence is underdetermined, even in principle, by the experiential facts. In other words, since there are at least two conflicting accounts of the nature of qualia, and yet each of these accounts is equally compatible with all the available evidence, including the totality of verbal and non-verbal dispositions associated with the supposed experiencing of qualia, it might be argued either that we are in some way radically mistaken about their true nature, or even, with

the eliminativist, that their very existence is unsupported by the evidence. Pursuing the most radical line, then, it might be argued that a certain degree of verificationism is justified in the case of experiential qualities, and that the availability of conflicting hypotheses regarding certain facts about qualia indicates that their existence is unverifiable in the required sense. Dennett cites a number of examples of experimental results which do appear to leave room, even in principle, for two conflicting qualia-based interpretations of a subject's visual experiences and associated behaviour. He begins optimistically in this vein with the comment that:

> A good way to understand a new theory [his own "Multiple Drafts" hypothesis] is to see how it handles a relatively simple phenomenon that defies explanation by the old theory [1991, p 114].

Here, we must assume in a verificationist spirit that for a phenomenon to "defy explanation" is for there to be two or more logically conflicting accounts of that phenomenon, each of which is compatible with the available evidence. As we saw in Chapter I, the eliminativist who is offering his thesis in support of physicalism is intent on denying the existence not just of experiential qualities [qualia] but of conscious experience in general. In the discussion that follows, then, although we shall refer to the hypothesis spurned by the eliminativist as the "qualia-based" hypothesis, it is to be borne in mind that at least some of the examples being considered are intended to show that even consciousness *per se* does not exist. The first example Dennett chooses is intended to undermine our conviction that qualia, in particular, are items of conscious experience; that there are two logically conflicting accounts of qualia each of which is compatible with all the available evidence.

<u>The Colour-Phi Phenomenon</u>.

Two small spots placed close together are alternately illuminated in rapid succession. To the observer, it is found that the appearance will be of a single spot moving from side to side. The apparently incongruous aspect of this phenomenon has two components. Firstly, although it is conceivable that the brain should "fill in" with a

moving image between the two spots when no such movement occurs in fact, it seems clear that, barring precognition, it can only do so *after* the second spot has been illuminated. Secondly, however, the experimental finding is that the subject is able to respond to the first spot, by pressing a response button, *before* the second has been illuminated. The problem is to provide a coherent account of how these findings can be reconciled. The observer reports an experiential sequence which he can only have constructed cognitively after a particular moment, while he in fact responds to the first spot before that moment.''

The problem is brought out more clearly in the version of the experiment conceived by Nelson Goodman [Goodman, p 85]. Here, the first spot is red and the second green. The experimental finding in this case was unexpected. It was that the moving spot appears to be red for the first half of its journey but green for the second half. The problem is that, barring precognition, the brain cannot know that the second spot will be green until it is illuminated. By this time, however, the illusory moving spot has completed its course and has therefore already changed colour in mid-course. It *seems* that the moving spot is experienced as undergoing a colour change prior to the green spot being illuminated, and it is difficult to see how this might be explained. However the facts are construed, there appears to be no logical explanation as to how a green spot *which has yet to appear* might create the impression of a colour change in the illusory moving spot. Nor does there seem to be any explanation as to how the spot can seem to be moving at all. For although he is able to register his awareness of the [apparently moving] red spot [by pressing a button] *before* the green spot is illuminated, the subject's memory of the spot includes information [it moving and turning green midway along its path] that could only have been acquired *after* the second spot has been illuminated and the direction of travel and change of colour thereby determined, and therefore after the subject has pressed the response button. How, then, are we to explain the fact that the subject's response occurs before he could possibly have become conscious of his moving red/green image?

---

11. See Kolers, P.A., and Grunau, M. [pp 329 - 335], for a detailed account of this experiment.

Dennett suggests that for anyone who holds that visual qualia are real attributes of conscious experience the most natural interpretation of the experience is to the effect that:

> your consciousness of the whole event must be delayed until after the green spot is (unconsciously?) perceived [p 115].

after which time the brain reconstructs the entire sequence of events and, so to speak, "presents it to consciousness" in the most plausible form. This suggestion he dubs the "Stalinesque hypothesis"; in essence, the hypothesis that all the information is withheld from consciousness until all the facts are in and a plausible story can be constructed to account for them. Thus, at the subconscious level, it might be supposed that each of the spots is observed in turn and only *then*, on the basis of the information gleaned from those observations, is the conscious image produced of a moving spot which changes colour midway along its travel.

Unfortunately, and this is where Dennett gets the opportunity to demonstrate the indeterminacy in the qualia-based account of what is going on, there appears to be good experimental evidence to show that this is simply not happening. Instructed to press a button as soon as the red spot is seen, the subject responds as quickly whether or not the green spot is subsequently illuminated. The implication is that there is no unusual delay in awareness of the red spot and certainly insufficient delay for even subconscious awareness of the green spot to arise prior to the subject initiating his response to the red spot [since the subject responds to the red spot *before* the green spot is even illuminated]. Hence, there can be no possibility of the subject's awareness of the green spot leading to delayed awareness of the red spot. Consequently, the Stalinesque hypothesis appears at least prima facie not to fit the facts. The evidence suggests that the subject is aware of [because he responds to] the stationary red spot even before the green spot has been illuminated. What the subject remembers after the sequence has been completed, however, is having pressed the button in response to the appearance of the moving red/green spot. He does not subsequently remember being conscious of a stationary red spot at all. At least, we might make this assumption here in order to give Dennett his best chance of establishing the presence of the indeterminacy he has in mind.

Given these assumed experimental findings, then, we are now in a position to consider the two possible accounts of the subject's conscious experiences. Briefly, either (i) he responds to subconscious awareness of the stationary red spot [presses the button] but his first *conscious* experience emerges later and incorporates the green spot, or (ii) he really is initially conscious of the stationary red spot [and perhaps even presses the button in response to that conscious experience] but subsequently forgets this experience in deference to the concocted memory of having experienced a moving red/green spot. The crucial question, then, is whether or not he *is* initially conscious of the stationary red spot.

1. The Stalinesque Version.

The first possibility, as Dennett observes [p 122], is that the subject responds to his *subconscious* awareness of the red spot in the timed experiment before he becomes consciously aware of it. If this is what happens, it remains possible that the brain does indeed become conscious of any of the events only after the green spot has appeared; it processes the incoming data about the entire sequence of events purely at a *subconscious* level, and only then is the final "invented history" [hence the title 'Stalinesque'] composed and presented for the first time to conscious awareness as a single moving spot. According to this account, then, the subject's first conscious experience is of a single moving spot which changes colour midway along its path, and he is correctly reporting this to be the case. For, on the assumption just that he does not remember experiencing the stationary red spot, it remains possible that he actually did not have that experience.

2. The Orwellian Version.

According to this version, the subject might *consciously* experience the stationary red spot *before* the green spot is illuminated, but then forget this experiential fact once the illumination of the green spot has been acknowledged, at least subconsciously, and his memory revised. The entire sequence will then always be remembered as a single moving spot which changes colour along its path, and in

that case the subject's report of his own experiences will be inaccurate. In this "Orwellian" version of conscious experience, then, the subject does indeed undergo experiences directly and in an unmisleading way [he initially experiences the stationary red spot without delay] but subsequent memorial distortions lead him to forget what he actually experienced. As a result of seeing the green spot, he concocts a revised memory of a moving red/green spot and forgets the original experience.

The problem according to Dennett is that from either the first-person or third-person perspective there is no way of deciding between the two accounts. What saves the Stalinesque version is the possibility of a *subconscious* button-pressing response to the red spot. From the timing involved it is clear that the response was initiated before the green spot appeared, but if the response was to subconscious cues it might still be true that the experience of the moving spot was the only [and correctly reported] *conscious* experiential sequence. On the other hand, the Orwellian version appears to fit the facts equally well, so long as we can accept the possibility that the subject's memory is not even reliable with regard to his own conscious experiences [we are assuming at least that he does not remember having experienced the stationary red spot]. And having allowed that there are two possible interpretations, each of which appears equally compatible with the reportable facts, there is, at least for Dennett, no conceivable reason for preferring one over the other. The suggestion seems to be that such a distinction is only possible on the basis of a fictional notion of experiential facts; the notion that there is a definite fact of the matter as to what the subject became consciously aware of and when, even though he is unable to remember that fact determinately.

Dennett's point in citing these two possible versions, then, is that since there is in principle no way of choosing between the two - no way of verifying one at the expense of the other - it makes no sense to hold on to the original assumption from which they were both generated. The problem supposedly shows up in the following way. In order to explain how the subject fails to remember seeing a single, stationary red light when he pressed the button, we have to say either that he really did see it [consciously] but the original memory was erased, or that he pressed the button in response to a

subconscious cue and really was not conscious of the red spot at the time. For Dennett there can never be any evidence to support a preference for either one of these hypotheses. But he thinks that if the assumption that there are particular items of experience leads to two mutually incompatible yet equally plausible hypotheses, then there must be something empirically unverifiable about the assumption itself. Consequently, he infers that there must be something wrong with *both* of the above interpretations.

> So, in spite of first appearances, there is really only a verbal difference between the two theories. ... The two theories tell exactly the same story except for where they place a mythical Great Divide, a point in time (and hence a place in space) whose fine-grained location is nothing that subjects can help them locate, and whose location is also neutral with regard to all other features of their theories. This is a difference that makes no difference. [p 125]
>
> If one wants to settle on some moment of processing in the brain as the moment of consciousness, this has to be arbitrary [p 126]

Now while there can be no doubt as to Dennett's motive for urging abandonment of this notion, it is by no means clear how he is supposed to be justifying such a move. For the fact that there are two possible explanations for a given set of reportable facts can hardly be construed as a reason for giving them *both up*. Dennett began by claiming that the phenomenon we have just examined "defies explanation by the old theory", but we have found no substantiation for this observation. The point is that once we allow that retrospectively memory can reshape the experiential facts there are two explanations available. The Orwellian version might be thought to stumble over the objection that a conscious experience which no-one remembers - "the way things actually, objectively seem to you even if they don't seem to seem that way to you" [p 132] - is a metaphysically dubious notion. The Stalinesque version, on the other hand, might be thought to challenge the intuition that the subject really would need to be conscious of a stimulus in order to respond to it by pressing a button. Whatever the reason for doubting either of these versions might be, however, there is no obvious reason for preferring Dennett's own account over the qualia-based account he is intent on undermining.

The Indeterminacy of Dennett's Multiple Drafts Account.

Once we look at Dennett's own account of events more closely we find that it too is underdetermined by the evidence in just the same way. Thus, consider firstly the sequence as described by the Stalinesque version of the qualia-based account. As we saw, this version depends crucially on the occurrence in general of a delay between the subconscious [button-pushing] and conscious [reporting and remembering] stages of response to a stimulus. Setting out the sequence of events in chronological order, then, the Stalinesque advocate would come up with following.

1. Red spot illuminated briefly.

2. Brain subconsciously registers 1.

3. Subject presses button to report 1.

[If there is too great a delay here the brain will consciously register 1 as a stationary red spot *and* subsequently remember having experienced it as such, and the example will not serve Dennett's purpose.]

4. Green spot illuminated briefly.

5. Brain subconsciously registers 4.

6. Brain forms account of overall sequence [single moving spot changing from red to green in mid-travel].

7. Brain becomes conscious of account in 6.

Remembering that the sequence just described is supposed to represent just one of two possible accounts of a sequence of events, however, consider now the following explanation of Dennett's own position. In response to Goodmans's observation that:

> ...the construction perceived as occurring between the two flashes is accomplished not earlier than the second [p 83].

he says that:

> The multiple Drafts model agrees with Goodman that retrospectively the brain creates the content (the judgement) that there was intervening motion, and this content is then

> available to govern activity and leave its mark on memory [p 128]

But:

> ...the brain doesn't actually have to go to the trouble of filling in anything with "construction" - for no one is looking [p 127]
>
> Why shouldn't the brain just *conclude* that there was intervening motion, and insert that retrospective conclusion into the processing stream? Isn't that enough? [p 128]

The Multiple Drafts Model differs essentially from the qualia-based account in that it denies the existence of qualia and consciousness in general. It claims that in the case we are considering, for example, information is being gathered and processed by the brain, and eventually some of this information is employed in producing the final edited version of what has been going on in the world and what information the subject has acquired. This final edited account is not presented to consciousness, however, but simply exhibited in the subject's consequent behavioural or dispositional states. Informational content can become "available to leave its mark on memory" [hereafter abbreviated to "*available for memorising*"] within the subject's brain, although this does not secure its place in the final draft and certainly does not entail the subject becoming *conscious* of that content. The important feature to notice about Dennett's treatment of the present example, however, is that it too is susceptible to both Stalinesque and Orwellian interpretations.

What would make it Stalinesque is the assumption that information about the stationary red spot is not available for memorising at any stage. For in that case the implication is that the subject responds [by pressing the button] to cues which are not yet available for memorising rather than to cues which are available but, in Orwellian style, would be subsequently rendered unavailable. Irrespective of whether Dennett is wielding Occam's razor in a responsible fashion here when he dismisses the consciousness-based options outright, however, the crucial point is that he could equally well have produced a revised version of his own multiple drafts model which would be essentially Orwellian in character.

Thus, according to the Orwellian version, the appearance of the red spot would become "available for memorising" *before* the button is pressed, and the subsequent memorial account would have this written out. There are therefore two versions of his own hypothesis each of which is compatible with all the evidence, and yet Dennett, far from calling into question the underlying assumption [about information becoming "available for memorising"] which generates these versions, has subscribed to one [the Stalinesque version] in preference to the other. He explicitly asserts that:

> There is no reality of conscious experience independent of the effects of various vehicles of content on subsequent action (and hence, of course, on memory). [p 132]

What makes this account explicitly Stalinesque is the insistence that whatever is meant by "conscious experience" there is no conscious experience which is not exhibited in behavioural or dispositional terms after the sequence has been completed [both the red and the green spot have been illuminated]. However he construes the subject's psychological condition when he presses the button, then, Dennett needs to face up to a real choice. Either the illumination of the red spot was "available for memorising" [or whatever he thinks is really going on when according to the qualia-based account he is "conscious" of the red light] or it was not. If it was, his explanation is Orwellian; if not, Stalinesque.

The point is that the uncertainty is not a consequence peculiar to the assumption of qualia or consciousness; availability for memorising has the same effect. As already observed, the essential problem, which opens up the possibility of rival accounts, is that the subject has no memory of having experienced [i.e., having consciously seen] the stationary red spot, even though he responds to it before the green light is illuminated. Once this much has been acknowledged, *any* explanation of events leading up to the memory actually retained [of the moving red/green spot] must be empirically underdetermined. Dennett's rejection of consciousness on the basis of this indeterminacy is grounded on his assumption that no conceivable experiment can show whether the subject became conscious of the initial image of the stationary red spot and then forgot it or was conscious of nothing at all until the one we know of emerged. But by the same token, exactly the same indeterminacy infects his own account. There will be no conceivable experiment either, in that

case, to show which possible version of his own account is correct.[12] The important conclusion in this case, therefore, is that in this respect Dennett's own hypothesis, according to which certain information becomes "available for memorising" at some point in the sequence, must be epistemically in the same underdetermined position as the traditional consciousness-based hypothesis he rejects.

Summary of the Indeterminacy Issue.

Referring back to the discussion in chapter I of the possible ways in which an observer might be mistaken about the nature of his conscious experiences, we saw that there were at least four distinct types of error to consider. The first was the error of making a false judgement about the physical state of affairs being observed [the straight stick looking bent in water, for example]. Quite clearly, both the qualia-based account and Dennett's alternative account of the physical sequence of events in the above experiment are susceptible to this type of error. For the agreed experimental findings, which need to be explained in terms of one account or another, indicate that the subject judges *wrongly* that he is observing a moving red/green spot and perhaps also that he presses the button in response to that moving spot. As we saw, however, the possibility of this sort of error in no way undermines the hypothesis that a conscious experience of some sort or other is occurring. The indeterminacy we are now considering is of this sort. It seems plausible to suggest that the subject might arrive at false judgements with regard to the events experienced, without the least doubt being cast thereby on the claim that he did, in fact, have conscious experiences with determinate characteristics.

---

12. Another possible objection to Dennett's rejection of qualia and consciousness is that there might, indeed, be enough *neurological* evidence to resolve the indeterminacy, at least in principle. Unfortunately, however, there is no obvious way of deciding which neurological state type [or token] constitutes *being conscious*, or being conscious of *qualia*. If we already knew that, the identity thesis would already have been vindicated; it has not, of course.

Secondly, however, the example shows that errors of the second type are also occurring. The subject can be wrong about the nature or characteristics of the experiences themselves. Thus, for example, we have to accept that according to the Orwellian version the subject has a certain experience [of the stationary red spot] but subsequently judges that he had no such experience. But again, this is not sufficient reason to conclude that there is no such phenomenon as experience, or no such qualities as qualia. In an earlier example it was perfectly plausible to suggest, for example, that the observer might wrongly judge his conscious experience to be of the reddish-orange quale, even though it is in fact a conscious experience of the red quale. As Robinson argues [1994, pp 195-8],

> What [Dennett] does not appear to allow is that there could be genuine phenomenology, in a traditional sense, but in which the phenomena are greatly affected by the kinds of conceptual activity associated with them [1994, p 195].

At least, we might want to insist at this stage just that there is genuine phenomenology in which "the phenomena [experiential qualities, etc.] are affected" in the sense that the subject's *judgements* about them might be mistaken in the ways already discussed. So we might be entitled to accommodate the indeterminacy exposed by Dennett by allowing that the subject might either make *incorrect* judgements [type 2 errors] about the phenomena experienced [in the Orwellian version] or *correct* judgements about the phenomena experienced [in the Stalinesque version]. Thus, for example, we might hold in a Stalinesque spirit that the first conscious experience of the subject really is, as judged, of the moving red/green phenomenal image. It is then only because we take it to be possible to form false judgements about one's phenomenal experiences [type 2 errors in chapter I] that we also allow that the the Orwellian hypothesis might nevertheless be the correct one. Hence the indeterminacy. By the same token, however, we would then have to concede that nor should the alternative account offered by Dennett be rejected on grounds of it being infected by this sort of indeterminacy. The fact, if it is a fact, that there is no conceivable experiment which would serve to establish whether or not the subject remembered, albeit fleetingly, the stationary red spot cannot be sufficient to justify the denial that events are ever remembered correctly, or even that there are any events to be

remembered. If we can allow that qualia might be experienced but misjudged by the subject we can allow also that fleetingly remembered experiences can be subsequently forgotten.

The point is, then, that Dennett's alternative [memory-based] account is indeterminate in just the same way as is the qualia-based account. But furthermore, there is no apparent reason to accept the indeterminacy in Dennett's account if in the light of the same indeterminacy the qualia-based account is deemed unacceptable. If Dennett's hypothesis is that at some point in time a particular piece of information [the illumination of the stationary red spot, for example] becomes "available for memorising", even though [as the Orwellian interpretation of events would have it] it might subsequently be rendered unavailable, then he is making a claim which cannot be verified and which therefore, presumably, suffers from unacceptable indeterminacy.

Dennett's Diabolical Operationalism.

The alternative is to reject *both* of the above theses in favour of the "the diabolical operationalism" according to which "what happened in consciousness is simply whatever you remember [presumably, after the completion of the entire sequence, in our example] to have happened" [p 132]. At certain stages in his exposition Dennett reaches the point of openly subscribing to this position.

> The Multiple Drafts model makes "writing it down" in memory criterial for consciousness. [p132]
>
> We might classify the Multiple Drafts model, then, as *first-person operationalism*, for it brusquely denies the possibility in principle of consciousness of a stimulus in the absence of the subject's belief in that consciousness [p 134].

Now, we might go along with Dennett's operationalist strictures for the sake of the argument and agree that we should not posit the existence of anything for which, even in principle, there can be no evidence. On that assumption, it seems clear that there would be no point in deliberating over the Orwellian/Stalinesque dilemma. Given

Dennett's assumption that there is no consciousness per se, but only the state of being disposed to make a judgement about what is going on in the world [call this state J], we might then agree that if there can never, even in principle, be any evidence to indicate that he has entered state J, no matter how briefly, with respect to the stationary red spot, it is pointless to speculate as to whether he in fact did enter state J on that occasion. Similarly, and by the same token, a *qualia-based* account of what is going on would be subject to the same ontological economy measures. There would be no point in speculating as to whether the subject really did have a fleeting phenomenal experience of the stationary red spot [call this being in state Q] if there can never, even in principle, be any evidence to indicate that he was in state Q on that occasion.

None of this need be regarded as controversial for present purposes. If the operationalist economy Dennett recommends is adopted, then any qualia-based account of experience will be forced to concede that the Orwellian/Stalinesque debate is meaningless and that only phenomenal experiences for which there can be some evidence should be acknowledged. Dennett applies this principle with scrupulous care to his own, judgement-based account; there is no point in wondering whether the subject ever entered state J with respect to the stationary red spot because there can never be any evidence that he did so. The problem is, however, that when he applies his operationalist economy to the *qualia-based* account, instead of insisting on acknowledging only those qualia or conscious experiences for which, at least in principle, there might be some evidence, he dismisses the qualia-based account altogether. If the Orwellian/Stalinesque debate is meaningless for qualia or conscious experiences, there just are no qualia or conscious experiences. Thus, he invites us to question the subject of the colour-phi experiment about what seemed to occur. The sort of reply Dennett would reject would be that:

> I know there wasn't *actually* a moving spot in the world... ... but I also know the spot *seemed* to move, so in addition to my judgement that the spot seemed to move, there is the event which my judgement is *about*: the seeming-to-move of the spot. There wasn't any real moving, so there has to be a real [phenomenal] seeming-to-move for my judgement to be about. [p 134]

But subsequently the real reason for Dennett's rejection of qualia and conscious experience altogether appears to have nothing to do with the indeterminacy so far discussed; it emerges in his explanation as to why he objects to this reply. He protests that

> "postulating a [phenomenal or experiential] 'real seeming' *in addition to* the judging ..... expressed in the subject's report is multiplying entities beyond necessity" [p 134].

There is a fundamental confusion in Dennett's dismissal of qualia and conscious experience here. The colour-phi experiment was cited as an example of a sequence of events to which a certain indeterminacy applies, either in terms of the qualia-based or judgement-based account of what states the subject entered. The resolution of the indeterminacy in either account is to be achieved by moving over to the ontologically more conservative operationalist model of events, according to which the indeterminacy disappears because only those states or experiences for which there can be evidence are acknowledged. But it is an entirely different matter to dismiss qualia or conscious experiences which have *not* been shown to suffer from indeterminacy in the chosen experiment. Thus, if despite the indeterminacy it still makes sense to retain a more economical version of the judgement-based account of the subject's states, it should also still make sense to retain a more economical version of the qualia-based account of the subject's experiences. Referring again to Robinson's comment,

> What [Dennett] does not appear to allow is that there could be genuine phenomenology, in a traditional sense, but in which the phenomena are greatly affected by the kinds of conceptual activity associated with them. [1994, p 195].

Dennett's operationalist economy will force us to take this more literally. For now, it is not just that we have to allow for the possibility of forming false [Orwellian] judgements or memories about the subject's experiences, but that we must deny even the *existence* of these indeterminate states or experiences. Thus, we end up literally conceding that the phenomena themselves are not as we first thought; questions about the subject's experience of the stationary red spot are unanswerable not just because of indeterminacy, but because there could, even in principle, have been no such experience. But even having conceded this much to

operationalism, there is still no reason for rejecting the qualia-based account of experience altogether, as we shall now see.

The Essentially Determinate Nature of Qualia.

The only possible way of salvaging Dennett's argument would be by showing that the qualia-based account alone is for some logically independent reason committed to items of conscious experience being determinate in the sense that Dennett shows them to be indeterminate. But it is a relatively simple matter to show that this need not be the case. In order to understand how this revised account of phenomenal experience might be formulated to Dennett's satisfaction, it will be helpful to quote his own account of the judgement-based multiple-drafts hypothesis at some length.

> Visual stimuli evoke trains of events in the cortex that gradually yield discriminations of greater and greater specificity. At different times and different places, various "decisions" or "judgements" are made; more literally, parts of the brain are caused to go into states that discriminate different features, e.g., first mere onset of stimulus, then location, then shape, later color (in a different pathway), later still (apparent) motion, and eventually object recognition. These localised discriminative states transmit effects to other places, contributing to further discriminations, and so forth ... ... The natural but naive question to ask is: Where does it all come together? the answer is: Nowhere. Some of these distributed contentful states soon die out, leaving no further traces. Others do leave traces, on subsequent verbal reports of experience and memory, on "semantic readiness" and other varieties of perceptual set, on emotional state, behavioral proclivities, and so forth. Some of these effects - for instance, influences on subsequent verbal reports - are at least symptomatic of consciousness. But there is no one place in the brain through which all these causal trains must pass in order to deposit their content "in consciousness".
>
> As soon as any such discrimination has been accomplished, it becomes available for eliciting some behavior, for instance a button push. ..... While some of the contents in these drafts will make their brief contributions and fade without further effect .... a few will even persist to the point of making their

> presence known through press releases in the form of verbal behavior [pp 134-5]

In addition, Dennett tells us that it is possible to "probe" this "skein of contents" [p 135] at different times with varying results. It is not made entirely clear how the probing is accomplished, but for the sake of argument let us assume simply that it is possible to "probe" by suddenly curtailing the process being observed [e.g. the alternating red and green spots flashing] at any particular moment and then looking for responses in the subject. So we might suppose that the colour-phi experiment is being conducted and the subject probed along the following lines.

Firstly, the entire sequence is simply allowed to run indefinitely and the subject asked to report on his experiences. As we have seen, he will report that a single light is moving from side to side and changing colour midway along its path as it does so. In response to the question whether he is able to see the red light as a stationary red light at any time he says that he is not.

Next, we decide to "probe" by suddenly discontinuing the sequence of flashing lights just after the red light has flashed, and before the green light has flashed. If there is to be any variation in response, we will have to assume that the subject will now report that his last experience was of a *stationary red light*.

How are these findings to be interpreted? In terms of Dennett's judgement-based account of experience it seems clear that the subject's disposition to judge that he sees a stationary red light has been created by the probing. In the first case he would never have entered the state of judging that there is a stationary red light, while in the second he was precipitated into the state of judging that there was. Dennett's operationalist interpretation of the findings might be, as already seen, that just insofar as the subject's disposition to judge has been altered his experience has been altered; there is no sense in speculating further as to whether or not the spots *seemed* to be one way or another independently of how the subject judged them to be. So if the sequence is just allowed to run the subject has one kind of experience, while if it is interrupted he has another.

But now we can imagine performing exactly the same experiment with a view to determining what sort of conscious *phenomenal* [qualia-based] experiences the subject is having. Again, if the flashing lights are allowed to run unchecked the subject [who is now, ex hypothesi, disposed to report his experiences in terms of the qualia-based account] will report that a single phenomenal image is moving from side-to-side and changing from red to green, or green to red, midway along its path. If the sequence is suddenly interrupted, however, he will report that his final experience was of a stationary red phenomenal image. So, barring other independent stipulations as to what counts as a phenomenal image, the natural operationalist conclusion would again be that what [conscious phenomenal] experience the subject actually has [and is disposed to report] depends on whether the sequence of events is interrupted.

In either version of experience, then, the operationalist economy has the effect of dissolving the indeterminacy. What is actually being experienced is simply what the subject is disposed to judge or report as being experienced. In either version, it is possible to alter the nature of the experience by probing at various points, but in neither case are we to infer that something was already being experienced independently of the subject's disposition to report it. In both versions, the continuous sequence of alternating spots was experienced as a moving single spot and that, according to the operationalist, is *all* that was experienced under those particular conditions. Probing the subject's brain provides evidence for other activity or information processing at a subconscious level; a fact which is equally compatible with each of the rival theses. The point we would now want to make is that if the detailed account of what is going on in the subject's brain [passage quoted from Dennett above] fails to undermine the operationalist principle in the case of Dennett's own judgement-based account of experience, there is no reason to suppose that it does not also do so in the alternative, qualia-based, account. If the subject's failure to report a stationary red light during the normal continuous sequence implies that he does not judge that that is what he sees, then his failure to report being *conscious* of a stationary red *phenomenal image* under those conditions has parallel, but qualia-based, implications. In other words, whether he actually experiences a stationary red quale or not is determined by whether or not the subject is probed.

As suggested earlier, the only apparent reason for objecting to this parallel treatment is that the qualia-based account insists that qualia, or conscious experiences in general, are supposed to be *essentially* determinate in a way which conflicts with the operationalist interpretation. But there is no prima facie reason why this should be so. Thus, the qualia-based account can readily accept everything Dennett has to say about the information-processing of the brain, as described in the multiple drafts model and summed up in the passage already quoted:

> As soon as any such discrimination has been accomplished, it becomes available for eliciting some behavior, for instance a button push. ..... While some of the contents in these drafts will make their brief contributions and fade without further effect .... a few will even persist to the point of making their presence known through press releases in the form of verbal behaviour [pp 134-5]

but then simply add that another upshot of all this processing can be the production of conscious experiences of qualia and other phenomenal properties. The fact that the nature of a subject's phenomenal experiences can be altered by "probing" [or interrupting the sequence] is then explained in exactly the same way as Dennett explains how the subject's judgements as to what is going on will change. Interrupting the alternating spots has the effect of probing or tapping into a different stream of informational processing, which leads to the judgement that there is a stationary red spot [and the production of consciousness in the subject of a stationary red phenomenal image]. Thus, according to this interpretation, the Stalinesque model is essentially correct, except insofar as it implies that conscious awareness of a sequence is in some sense artificially delayed in order to accommodate the green light in the final conscious draft. We might suppose that the emergence of processed information into consciousness in general takes longer than the temporal spacing between the red and green spots being illuminated. The determinate, operationalist account of *qualia* and consciousness can then be summed up as follows.

Firstly, if the spots are allowed to alternate without interruption, the subject responds to the brief illumination of the red spot subconsciously [i.e., without having become conscious of a

phenomenal image of the spot]. Then the green spot is briefly illuminated and, still before any phenomenal content at all has emerged into consciousness [because the spots alternate quickly enough], the brain processes all the material available and decides on the model according to which a single moving spot changes from red to green midway along its path. A conscious phenomenal image based on that model is then produced. The subject becomes conscious of seeing such a spot. The alternative possibility is that Dennett probes the subject's brain by interrupting the sequence as before. As we have seen, this has the effect of tapping into a different stream of informational processing which leads to the subject forming the judgement that he has seen a stationary red spot. We need only add that according to the qualia-based account this stream also leads to the production of a corresponding phenomenal image in consciousness. Hence, the experimental findings of the colour-phi experiment have been accommodated without the need to infer that there is indeterminacy in the qualia-based account. The multiple drafts model of information processing in the brain turns out to be fully compatible with the existence of determinate qualia.

Summary of the Indeterminacy/Operationalism Debate.

In short, what Dennett will have to concede is that he is not in a position to reject the qualia-based account on grounds of indeterminacy. If Dennett's "diabolical operationalism" is his own way of resolving the indeterminacy we have just exposed in his own account, by rendering "memory criterial for consciousness", then the indeterminacy in the qualia-based account can be resolved in a similar way. Thus, if it is to be ordained that the indeterminacy is intolerable, the qualia-based account can be preserved by simply conceding that it is only permissible to acknowledge experiential qualities or characteristics which the subject is able to remember. There can be no objection in principle to conscious experiential qualities [qualia] for whose existence memory is criterial. In terms of either account, then, the resolution of the indeterminacy is essentially Stalinesque [in that only what is remembered is acknowledged at all]. If, on the other hand, the diabolical operationalism is deemed intolerably coarse, it must be so for the two competing theses alike, and the idea that an observer might consciously experience a stationary red spot and then forget that he

did so is reinstated as a respectable [and Orwellian] hypothesis, along with the indeterminacy it was cited to expose.

In the final analysis it is important to bear in mind the ultimate purpose of the current enquiry. What we are exploring is the possibility that phenomenal items such as qualia might *seem* to exist even though they do not. Hence, making sense of this possibility necessarily involves the consideration of qualia which do at least seem to exist; the subject is at least disposed to report that he experiences them. But we have found no compelling reason for insisting on a notion of qualia which exist even if the subject fails to notice them. Thus, in particular, the indeterminacy introduced with the Orwellian interpretation of events involves supposed qualia which do not even *seem* to exist, and therefore whose claimed non-existence is unlikely to be controversial. What Dennett really needs to explain is what is going on when qualia *do* seem to exist, and to do this convincingly he must cite cases in which the subject is at least disposed to judge that they do exist, and then go on both to show that they do not exist and to explain how our false conviction that they do might have arisen.

Another Example; Dennett's Beer-Drinker.

Other examples cited by Dennett confirm our findings. When people first experiment with the taste of beer, he suggests, they often find it distinctly unpalatable. With perseverance, however, there is often a change of heart on this point. Now the question posed for the qualia-based account is whether the experienced beer-drinker begins to experience a new taste-quale from beer or merely becomes accustomed to the original quale. Dennett thinks it is impossible to say. In other words, he claims that there is irresolvable indeterminacy as to whether he remembers the original quale accurately or not. And if, as before, there are two conflicting hypotheses about qualia each of which is compatible with all possible evidence, Dennett's response is simply to deny that there *are* any qualia at all.

> So if a beer drinker furrows his brow and gets a deadly serious expression on his face and says that what he is referring to is "the way the beer tastes to me right now", he is definitely kidding himself if he thinks he can

> thereby refer to a quale of his acquaintance, a subjective state that is independent of his changing reactive attitudes. It may seem to him that he can, but he can't. [Dennett p 396]

He provides no explicit argument in support of his eliminativist claim in this context. He merely deems it impossible to tell whether our "beer quale" changes with experience or our reactive dispositions to the original beer quale do the changing over time. Now we might even want to concede to Dennett what seems to be a fact; namely, that to a certain extent our memory of the original beer quale will not be sufficiently reliable to enable us to decide how much of the adaptation was due to a change in the original quale and how much was due to a change of reactive dispositions to a given quale. We might even concede further that there is, even in principle, no way at all of determining which of the rival hypotheses about remembered experiences is true. But even so we would not thereby be relinquishing our claim to qualia realism. If we were, it is difficult to see how parallel considerations might not also infect Dennett's realism with regard to the state of being "available for memorising".

Turning the point around, suppose that qualia do exist, so that there is a fact, to which Smith has reliable access at the time, regarding the specific qualitative experience which he is currently undergoing. Given that assumption, it is then clear that an unreliable or misleading memory would be sufficient cause to wonder about the objectivity of Smith's comparison of the character of his beer quale at different times. In response to Dennett's claim that the beer drinker is mistaken even about the character of his *contemporary* beer quale, however, we would insist that this is an entirely separate matter which his [memory-based] indeterminacy has singularly failed to infect. An example of an incorrect judgement of this sort would emerge if Smith were to assess his "pinkness quale" pP, for example, as being homogeneous when in fact it is not homogeneous, or the quale Rp which he experiences on looking at a ripe tomato as being just like the quale Gp which he experiences on looking at the leaves on the tomato plant, when in fact they are different.

Take the lifelong beer drinker. If it seems to him that the beer quale has remained constant over the years but his liking for that quale has only developed recently, then that is how it seems to him. If, on the other hand, it seems to him that the beer quale itself has changed over time and that he likes the way it has ended up more than the way it began, then that is how it seems to him. Either of these judgements is dependent on Smith's memory and hence susceptible to its unreliable nature. However, nothing in the argument so far has shown that there are no clear facts about the phenomenal character of his experiences *at this moment*. In order to be mistaken about these facts, Smith would have to arrive at a belief or judgement about the character of his present seeming phenomenology which is false, and which does not constitute a memory-dependent comparison with previous qualia. Although his beer quale has such-and-such a character now, he would judge that it does not in fact have that character. Once again, then, the unreliability of memory or even the indeterminacy of remembered facts has no obvious repercussions for qualia-realism. Suppose that at the end of the comparative process it seems to him that his beer quale has remained constant over time. It seems so to him because his beer quale is now a particular way and he remembers [perhaps falsely] that his beer quale of twenty years ago used to be that way also. It is in this *comparison* that the uncertainty lies. Nowhere in the argument have we had reason to doubt, with Dennett, that he is capable of referring accurately to "the way beer tastes to me right now", or that there is no such property as a current beer-quale.

Indeterminacy in Cases Which Do Not Involve Memory.

A similar defence strategy can be readily devised against counterinstances which do not depend on memory. On looking at a red spot, we might initially claim that Smith's experience of the spot has a qualitative character, or quale, Rp and that this unique character is immediately and unmisleadingly available to Smith's introspection. In accordance with Dennett's operationalist principle, whatever the subject seems to experience is what he does experience. On that assumption, it follows that there is a sense in which he cannot be wrong about that character; that is to say, Rp cannot seem to Smith, introspectively, to be other than it in fact is. Suppose, then, that when a red spot is placed next to a blue

spot, the effect of contrast results in the original character, Rp, of his experience of the red spot *seeming* to change slightly. The subject is now disposed to judge that he is experiencing the reddish-orange quale ROp. We can immediately see that, unless the indeterminacy is resolved by a move towards operationalism, there are two competing explanations for this phenomenon.

The first is that the experiential quality Rp is indeed immediately available to Smith but during the experiment is replaced with a slightly different yet equally available quality ROp. According to this hypothesis, then, Smith's experiential quality does not merely seem to change; it really does change and Smith is under no illusion with regard to that change. He is misled, not about the quality of the experience but about the objective colour of the spot. He correctly judges that it looks reddish-orange [produces the quale ROp normally produced by reddish-orange objects in standard conditions] but would be wrong to infer that it is *in fact* a reddish orange spot [i.e., that in standard conditions it produces the quale ROp]. Clearly, this first explanation is compatible with the existence of colour qualia and, indeed, with Dennett's operationalism, since at no time is the character of the phenomenal experience claimed to be other than Smith judges it to be. There is, however, an alternative explanation.

Thus, it might be argued that although the spot seems to Smith to produce the reddish-orange quale this is because, although he is actually experiencing Rp throughout the experiment, the proximity of the blue spot leads him to *judge wrongly* that he is experiencing ROp. As in the colour-phi experiment, then, there are two distinct hypotheses about Smith's experiential qualia. Either he experiences a change from Rp to ROp and judges this correctly or he experiences Rp throughout but judges incorrectly that it changes to ROp. In other words, since he experiences a change of some sort, that change can involve either the quale itself or just his judgement of that quale. Furthermore it seems that there is no conceivable evidence which would settle the matter one way or the other. For once we accept the fallibility of his judgement with regard to the qualitative nature of his own experiences the indeterminacy noted earlier emerges once again.

The parallel with the previous examples should now be apparent. Dennett's operationalist approach would incline him to suggest that if there are two competing but indeterminate hypotheses about qualia the reasonable course would be to abandon talk of qualia altogether. But again we would strongly resist this move. What he would say in this example is presumably that the distinction between our two qualia-based hypotheses is a distinction without a difference and that the origin of it, the notion of qualia itself, should be abandoned. We would then be left with the eliminative thesis that the subject makes judgements about the colour of the red spot, but that there are no phenomenal properties or qualia to help him make those judgements. The subject clearly believes that the red spot seemed to change colour, and insofar as he believes that this is the case it *is* the case. Over and above the fact that the red spot *seemed* to turn reddish-orange when brought next to the blue, however, there are no experiential, or phenomenal, facts to describe.

This example is particularly interesting in the present context because it brings out the crucial difference between the rival hypotheses. Once we allow the possibility in principle that the subject might judge a quale to be other than it is we have two competing hypotheses, as in the previous case involving memory. Just as memory must be conceded as fallible, so too must judgement. This assumption renders plausible the second interpretation in which the subject experiences Rp and yet wrongly judges it to be ROp. Admittedly, Dennett might insist that it is impossible to make sense of a notion of experiential qualities which allows for [type 2] errors of judgement and thereby opens the way for the second qualia-based interpretation.

Even if we concede, however, that in accordance with Dennett's operationalism the latter is impossible - that a quale being a certain way precludes the subject judging it to be another - we are still left with an intelligible qualia-based version of Dennett's operationalist account. For according to the first interpretation, the subject begins by experiencing Rp as a result of looking at the isolated red spot and ends up experiencing ROp when the blue spot is illuminated. Since he correctly judges the quality of his experience to have changed in just this way there is no conflict in this case between how the experience really is and how it is judged to be.

There is therefore no need to consider whether it is possible to judge an experiential quality to be other than it is - whether we need to make sense of a spot *seeming* to be red but seeming to seem to be reddish orange [from the reference to Smullyan 1981, Dennett, p 132]. So once again Dennett's insistence on removing indeterminacy by restricting the facts to those reported by the subject carries no implications for consciousness or qualia per se.

The point is that Dennett's operationalist account of *judgement* as an infallible indicator of experiential fact can be divided into two distinct parts. In the first place it entails that only one qualia-based hypothesis is available [i.e., the first]. If it is impossible for Smith's experience to be one way and be judged another, then, we must assume that if he experiences Rp when looking at the single red spot, but judges or believes that a qualitative change in his experience is brought about by the blue spot being produced, then he must be correct. Judging that the experiential quality changes from Rp to ROp entails that it does change from Rp to ROp. There is no significant conflict here with the traditional Cartesian view that experiential qualities are immediately available to the subject's introspection. The second inference urged by Dennett, however, is more controversial. For even construing Smith's judgement thus, as a reliable indicator of the qualitative experiential facts, there is still no inclination to agree with Dennett that the judgement is *constitutive* of those experiential facts. None of the examples or arguments so far considered suggests in any way that phenomenal properties of experience *seem* to occur but in fact do not.

Conclusion.

The approach we have been considering was intended to establish that there is something incongruous or indeterminate about the positing of qualia - that there are known facts about a subject's colour vision, for example, which it is difficult or impossible to explain on the assumption that he experiences qualia. What we have found, however, is that the known facts present no such difficulties. The most that can be said in support of Dennett's memory-based counterexamples, for instance, is that if the positing of qualia implies that the subject becomes conscious of his experience of a stimulus at a specific moment, then the subject's memory evidently

falls short of recording such facts fully and faithfully. But this leaves available at least two possible ways of salvaging qualia. Either the indeterminacy engendered by a necessarily fallible memory is to be tolerated, along with the qualia whose existence it was cited to refute, or the indeterminacy is resolved by adopting a more instrumentalist notion of experiential qualities. According to the first option, it is legitimate to conclude that if the subject does experience qualia he both loses access to facts about them [the fact as to whether or not he experienced a stationary red spot in the colour-phi experiment, for example] and even possibly that he creates objectively false memories of them [remembers that he became conscious of them at a time when it was impossible]. But exactly the same objections can be levelled against Dennett's own hypothesis; that what we suppose to be the subject's conscious awareness of stimuli [and hence qualia] is just the availability of stimulus information for memorising, judging, or whatever. For in that case the subject's memory loses facts about what is or is not available for memorising, and introduces objectively false memories, in just the same way. If the indeterminacy is deemed intolerable and the instrumentalist option adopted, however, there is still no obvious reason for preferring Dennett's judgement-based account over the qualia-based account of experience. For if, in the case of judgements or memories, we are only to count as having been experienced that which the subject is disposed to report, then we have seen no reason why the same criterion should not be applied to consciousness and qualia *per se*.

If the eliminativist is to succeed in his attempt to establish that the qualia-based hypothesis claims the existence of items which do not exist, then, he must presumably appeal to the unverifiability of this claim on other grounds. For eliminativists in general, there must be just insufficient evidence to warrant the positing of experiential qualities in addition to the facts about the subject's judgemental dispositions or neural states. And it is precisely because the positing of such qualities is not prompted by the *physico-dispositional* facts that the latter fail to afford the appropriate evidence. For the eliminativist a subject might be in a particular *neural* state which prompts him to judge that he is seeing red, for example, but need not also recognise that he is in a particular *qualitative* state as of seeing red in order to do so. Thus, we are not entitled simply to assume with Strawson that:

> None of the oddities and indeterminacies of experience detailed in *Consciousness Explained*, for example, so much as touch the validity of our basic grasp of the nature of the experiential... [Strawson, Galen. pp 99-100]

Specifically, once the eliminativist's claim is seen to be simply that the positing of consciousness and qualia is *redundant* [has no additional explanatory power], rather than that it is incompatible with, or underdetermined by, the evidence, it would be unreasonable not even to consider his proposal. To reject the whole project as "irrational and unscientific" on the sole ground that "the existence of phenomenological features of mental life is one of the most obvious and unavoidable categories of data with which we are presented" [Strawson 1994, p 67] is to *presuppose* that the belief is sufficient evidence for the fact. For, in offering an alternative explanation for the belief according to which the positing of consciousness and qualia is redundant, it is precisely this presupposition which the eliminativist sets out to challenge. The legitimate procedure at this point would be to look for evidence that the eliminativist's "redundancy thesis" is mistaken; that there is, indeed, evidence for which the better explanation would be in terms of consciousness and qualia. This theme will be taken up in subsequent chapters.

There is, however, a preliminary question which deserves attention. That is, does the eliminativist have any intelligible way of distinguishing between his own position, that qualia and experience are non-occurrent phenomena, and the apparently distinct [reductive] thesis that they are in fact physico-dispositional in constitution and character? From the foregoing considerations it seems clear that the distinction is at least intelligible in the case of Yetis and bears, where the identifying characteristics for Yetis can be specified in physico-dispositional terms. The case of qualia and experience is more problematic, however, since here there are, ex hypothesi, no such terms available in which to draw an intelligible distinction. Secondly, if he is able to distinguish his position intelligibly from reductivism, can he *justify* his thesis given the evidence available to him? It is to a thorough treatment of these questions that we turn in chapter III.

Chapter III

THE ELIMINATIVIST/REDUCTIVIST DISTINCTION.

Introduction.

In this chapter we shall explore the possibility of distinguishing, intelligibly, between two forms of physicalism already characterised as "Qualia Eliminativism" [QE] and "Qualia Reductivism" [QR]. Initially, for the sake of simplicity, the respective theses will be characterised as:

[QE] Qualia do not occur.

[QR] Qualia are occurrent physical properties.

Some initial clarification of these positions is in order.

As explained in the introduction, we shall regard properties as universals; the assumption is that token sensory experiences are to be construed as of a particular *type* just if they *have*[13] an instance of a particular property. An occurrent [token] experiential state [for example, Smith's state S at time t] will then be referred to as a token state of a particular type [H], or [headache], just if it has an instance of the property H.

Also, we assume that the term [S] is used by any particular individual to refer to the set of physical *items* [states, properties of states and events] which he believes to occur; *to have occurred* at one time or another. Thus, the contents of [S] will vary from one individual to another depending on his beliefs, or ontic commitments. Initially, we shall assume that, logically prior to the debate about qualia, QE and QR *agree* on the contents of [S]. We shall assume also that the items in [S] are all *physical* items, the

---

13. We remain idiomatically non-committal here on the issue of whether a token experience *possesses*, *exhibits* or *provides introspective access*, to an instance of a property.

items recognised by the physicalist as occurrent, even if they can only be fully *characterised* by invoking dispositional properties.

The items they already accept as belonging to [S] might then be referred to as the "paradigmatically physico-dispositional" [PPD] items; occurrent items which are already accepted as being:

> (i) Physical in nature [as tentatively explained in the introduction].
>
> (ii) Physico-dispositional in character; the item will be of a particular neural type [N] if it exhibits an instance of a neurally characterised property N, for example, and of a particular dispositional type [D] if it exhibits an instance of the dispositionally characterised property D.[14]

In the case of *properties*, then, a property will be said to be a member of [S]; that is, a PPD property, just if there have been occurrent instances of that property and the property itself is a purely physico-dispositional characteristic.

Having made these initial assumptions, then, we are now in a position to redraft the two theses as:

> [QE] Qualia are not PPD properties (properties belonging to [S]) and do not occur.
>
> [QR] Qualia are PPD properties (properties belonging to [S]) and [therefore] do occur.

And from this we can see already that we are likely to encounter a fundamental problem with regard to distinguishing the two theses.

---

14. Items can have both physical and dispositional properties, and therefore be said to be of physical types [having physical properties] and dispositional types [having dispositional properties], at the same time. We do not even need to presume that there is an intelligible distinction to be drawn between the two types of property for present purposes. The only assumption is that even dispositional properties are *physical*, in the sense explained in the text.

The problem is this. If both QE and QR are physicalists in the sense outlined in the introduction, they will agree that all occurrent items are PPD items; in our version of physicalism, items which are intelligible and epistemically available from within the third-person conceptual framework and perspective of physical theory. As we saw, however, the physicalist purports to include all occurrent *properties*, or, more accurately, all occurrent instances of properties, in his account of the world. So we might suppose initially that where the two theorists disagree crucially is over *which* [intentionally inexistent] properties have occurrent instances in [S]. But QR does not claim that qualia are properties *in addition* to those in [S]; rather, that they are some of the properties already included in [S]. The items referred to by [QD] as occurrent instances of "qualia" are in fact occurrent instances of PPD properties. But this implies that any two physicalists who share a common ontic commitment just to the members of [S] will deny the occurrence of any other properties.[15] The difference of opinion between [QE] and [QR], therefore, appears *not* to be an ontological one.

What, then, might the difference of opinion amount to?. If, ex hypothesi, the two theorists are agreed as to which properties have occurrent instances [as members of [S], we must suppose that the dispute amounts to a disagreement over the way in which the "qualia" acknowledged by QR should be described; the properties or characteristics which can be legitimately ascribed to them as occurrent instances of PPD properties. Thus, QR might construe QD's

---

15. There is a further possible position to the effect that [S] might be *expanded* to incorporate qualia as additional members. In principle there are two distinct ways of doing this. Firstly, it might be proposed that qualia *can* be physico-dispositionally characterised, but have yet to occur. If they should occur in future, they will be members of [S]. Alternatively, qualia might be incapable of physico-dispositional characterisation, and yet occur as additional members of [S]. The first proposal will be explored later in the chapter, while the second will be disregarded as representing a philosophical position which goes beyond the scope of *reductive* physicalism.

description of qualia as a correct description of PPD properties which occur in [S], while QE might deny that any occurrent properties in [S] satisfy that description. It is not at all clear that this sort of disagreement is not an ontological disagreement, however. For we might reasonably say that whereas QR accepts the occurrence of qualia as properties bearing certain characteristics, QE does not. This seems to be an ontological dispute over which *properties* have occurrent instances in [S].

I propose to sidestep the question of whether the disagreement between QE and QR is ontological or descriptive. What interests us here is whether they can even intelligibly specify the same property in their respective theses. Even if they can, however, we need not be detained by the question of whether their disagreement is ontological [i.e., whether one claims that certain properties occur and the other does not] or descriptive [i.e., one claims that occurrent headaches have certain properties and the other does not]. For the important difference between the rival theorists is how they describe occurrent headaches. The answer to this question is to be sought in the properties or characteristics ascribed by each to occurrent headaches. If QR says that occurrent headaches have a certain property X, for example, we can understand that for him occurrent headaches are of type-[X]. Whether this is to be deemed a statement of his ontological or descriptive commitments seems relatively unimportant.[16]

The Physicalist's Account of Introspection.

As indicated briefly in the introduction, it seems indisputable that

---

16. We considered Rorty's [unanswerable?] question of how wrong we can permit someone to be about an item [e.g., a bear] and yet still take them to be referring to a bear. In the present case, what we need to establish is whether the *properties* referred to by QE and QR as "qualia", and ascribed only by QR to headaches, are one and the same. Irrespective of whether the dispute over the occurrence of "headaches" is to be construed as an ontological dispute, it only has any meaning if the same "qualia" are being referred to by each.

when Smith determines in the first-person perspective introspectively] that he has a headache, he is at least recognising that *something* is occurring. Furthermore, it seems clear that he is able to distinguish introspectively between headaches and say, bouts of nausea. In accordance with common sense, then, we shall assume that at least some recognitional and discriminatory ability is in evidence in such cases, and therefore that there is no difficulty for either in determining that they have a headache. Before attempting to differentiate between QE's account and QR's account of what is thereby being recognised and discriminated, however, it will be instructive to explore the facts on which, as physicalists, they would agree. Consider the case of colour perception and discrimination.

Firstly, they are agreed that there are no qualia *in addition* to the occurrent PPD properties. As human beings we are able to recognise and discriminate between the various shades of red, for example, and what Dennett thinks is going on when we do so is just that:

> When we make these comparisons 'in our mind's eyes', what happens according to my view? Something strictly analogous to what would happen in a machine - a robot - that could also make such comparisons. .... Suppose we put a color picture of Santa Claus in front of it and ask it whether the red in the picture is deeper than the red of the American flag (something it has already stored in its memory). This is what it would do: retrieve its representation of [the American flag] from memory, and locate the red stripes (they are labeled "red #163" in its diagram). It would then compare this red to the red of the Santa Claus suit in the picture in front of its camera, which happens to be transduced by its color graphics system as red #172. It would compare the two reds by subtracting 163 from 172 and getting 9, which it would interpret, let's say, as showing that Santa Claus red seems somewhat deeper and richer (to it) than American flag red. [Dennett, 1991 p 374]

In addition to claiming that the imagined robot is capable of making just the same colour discriminations as we are, then, Dennett is insisting that since no qualia feature in the case of the robot, nor do they in our own case.

> The [robot] certainly doesn't have any qualia, so it does indeed follow from my comparison that I am claiming that we don't have qualia either. [pp 374-5]

Like QR, he believes that although we are able to discriminate between, for example, the various shades of red, we actually do so purely in virtue of the PPD states associated with colour perception and discrimination. If a sophisticated, but physically constituted, robot is able to distinguish between two reds as in Dennett's example, and yet do so in virtue purely of PPD states, properties, etc., then it follows that according to *both* QE and QR it would seem plausible to describe the robot as having "judged" [in much the same sense as a thermostat might be loosely described as "judging" that the temperature is too high] that one red is deeper than the other. But, of course, the robot which QE and QR envisage need not be assumed to "judge" or "believe" in any sense that non-PPD properties of any sort whatever are involved in the process; its judgements might be about purely PPD properties [e.g., temperature]. Whereas we as humans *might* judge that seeing the red of the Santa Claus suit produces a particular non-PPD quality, or quale, the mechanical colour-discriminator might simply judge that it is seeing red #172, or that the red it is seeing is deeper and richer [has a higher index number] than the red of the American flag. Similarly, if the robot is more sophisticated, we might imagine that it has the further ability to judge that it is in a particular discriminative state. Whenever it judges that the room is too warm, for example, it will exhibit its ability to discriminate temperature by turning off the heat. But it might additionally exhibit a mechanical equivalent of self-consciousness in the form of an ability to register the fact that it is in the state of recognising that the room is too warm by, for example, turning on a red light. Even so, the robot might not be making any judgements about non-PPD properties or states.

According to both QE and QR, we are like the robot in that our seeing an objective sample of red #172 does not involve the recognition or discrimination of any non-PPD properties, since we can at least affirm that their shared ontic commitment excludes any non-PPD properties whatever. Instead, seeing red #172 amounts just to being in an appropriate PPD state R172. Looking at a Santa Claus suit in standard conditions, Smith sees red #172 [acquires state R172] which disposes him to judge that the colour of the suit is red

#172]. Being in R172 usually also gives rise to another state, S172, which disposes him to say such things as "Ah, yes, now I am seeing red #172", or "Now I am in state R172". But a sophisticated robot might also be equipped with equivalent self-monitoring faculties.

For the sake of completeness, we might finally imagine a robot which responds to seeing red by not only acquiring states R172 and S172, but in addition gets into a state T172 in which it *reports* that it is experiencing a non-PPD quale Q172. It might thus be assumed to be identical with the qualia-*dualist* in respect of the PPD states it exhibits. The physicalist's response to this additional disposition would amount to the claim that although the robot undeniably has the disposition to *report* the experiencing of qualia, nothing is actually occurring which cannot be fully accounted for in PPD terms. Thus, in eliminative style he might insist that like the robot the qualia dualist is referring to no occurrent properties, or in reductive style that he is referring, perhaps misleadingly, to occurrent PPD properties which belong to [S].

We can now return to the case of Smith's headache. The common physicalistic position of QE and QR here would be that although Smith evidently is able to discern and discriminate various of his bodily states [state-types; states bearing instances of particular properties], whatever he discerns and discriminates are all PPD states and properties contained in [S]. So, for example, the headache Smith discerns at time t must be a token PPD state characterised as having some PPD property H in virtue of which it counts as being a headache. Given that the two physicalists would agree to this extent, then, we can now explore possible respects in which they might *disagree* over their respective descriptions of what such a headache amounts to, and in virtue of which their respective positions might be intelligibly distinct.

If, as an eliminativist, Dennett disagrees with QR at all in respect of which occurrent properties [qualia] are discerned in introspection, we can assume that he does so in virtue of QR's ascription to those qualia of some characteristic X, which QE does not believe to occur. In other words, when QE says that such qualia do not occur, he can be interpreted as saying at least that properties with characteristic X, or type-[X] properties, cannot be ascribed to the headaches discerned in introspection.

The [QE]/[QR] Distinction.

In order even to distinguish his position intelligibly from [QR], then, QE will need to be able to cite some intelligible fact about QR's "qualia" which ensures that each is referring to the same qualia. He must establish that the qualia referred to by each are properties of some intelligible and specifiable type-[X] such that the bearing of property X is sufficient for an item to be a quale.[17] Once this much has been achieved it will become meaningful to consider the rival theses concerning the ascription of qualia to occurrent introspectible headaches, for example.

We can begin to consider what property X might be by recalling the redundancy thesis to which the eliminativist was shown to be committed in chapter II. Thus, despite the efforts of such eliminativists as Dennett to demonstrate that in the human case the positing of irreducible consciousness and qualia in some way *conflicts* with the physico-dispositional facts, we argued in chapter II that this is simply not the case. We argued that an account of colour perception and discrimination which resorts to non-PPD properties of some kind seemed to be no more or less compatible than the physicalist's alternative with the occurrent physico-dispositional evidence. Hence, the strongest position available to the physicalist [that is, if we assume with the physicalist that the PPD account of colour perception is complete] is the claim that qualia-discourse is simply *redundant*. Thus, we might state the redundancy thesis as:

> [RT] There is no PPD evidence which would justify the positing of occurrent but irreducibly non-PPD properties [qualia, for QD]. However, there is nothing about qualia-discourse which conflicts *logically* with the PPD evidence.

---

17. This appeal to the properties of properties need not be problematic. Thus, Redness and Blueness are different properties, but they share the common *property* of *Being colour properties*. QE must cite a property X of QR's qualia which no occurrent introspectible properties have, and which therefore precludes the physicalistic reduction of qualia to members of [S].

The eliminativist's position depends crucially on this claim since, if there were any such evidence, he would not be in a position to substantiate his rejection of [QD]. Thus, if QD claims that he is able to discriminate colours [i.e., exhibit his PPD traits] by virtue of the qualia they produce, the eliminativist will have to reply that colour discrimination can be given a purely PPD account and that there is no evidence for the qualia to which QD refers [that is, qualia which are irreducibly non-PPD properties]. But we have seen that there is no compelling reason to suppose that QD's claim conflicts logically with the PPD evidence. QD can accept the PPD account in full and simply add that it is accompanied by the experience of irreducibly non-PPD qualia.

In the present context, however, the eliminativist cannot cite [RT] in an attempt to repudiate [QR], since both QE *and* QR believe that the PPD states R172 and S172 [in our simplified account], and additionally T172 in the case of the qualia-dualist, are sufficient to enable the subject to display all the red #172-related dispositions and capabilities he in fact does display, and that there are no further facts to be accommodated by the positing of the occurrence of some *additional*, non-PPD property, Q172. The dispositions and capabilities which the recognition of that property is supposed to make possible are already made possible by the acknowledged PPD properties and states included in [S]. Hence, the positing of any additional properties which would make those dispositions possible is redundant both for QE *and* for QR. The crucial difference between the two positions appears to be that whereas QR claims that his "qualia" are just some of the PPD properties already mentioned, QE claims that QR's "qualia", specified determinately as properties of type-[X], do not even occur. In other words, he must claim that none of the PPD properties which *can* be ascribed to headaches has characteristic X.

The present task for QE, then, is to cite some property X which he can determinately ascribe to QR's qualia. As a first attempt, he might suggest that the property X which can be ascribed to QR's qualia is not even *intelligible*. If he can substantiate this allegation, he will then be in a position to reject QR's qualia as being [intentionally inexistent] properties of an unintelligible type.

If QR claims that there are qualia and that they are reducible to PPD properties in [S], QE's response might be that, while he understands the proposition that there are occurrent properties which are reducible to PPD properties, he simply fails to understand *which* properties are being referred to. Thus, QR's reference to qualia *in particular* as reducible properties is simply unintelligible. The problem with this response, however, is that it seems clear that the characteristics of qualia to which QR seems to be committed just *are* intelligible. QR is likely to claim, for example, that there are certain epistemic facts about qualia in terms of which it makes sense to say that qualia are reducible to PPD properties. In particular, he might hold that the *diagnostic* epistemic characteristic of qualia is just that the fact that they are PPD properties can be ascertained only *a posteriori*.

Taking this initially to be characteristic X, then, QE can only establish that QR's qualia do not *occur* if he can establish that in the epistemic situation Eq in which QR claims to be able to pick out a quale Q, but not to be able to ascertain that it is a PPD property, the *occurrent* properties which can be picked out epistemically *can* be known to be PPD properties.[18] Thus, he must at least claim that, in Eq the fact that each occurrent property p is a PPD property can be determined a priori, by logical or conceptual inference from the facts known in Eq [If a posteriori investigation were required for this determination, further facts would be incurred and the epistemic situation would no longer be Eq, but some other epistemic situation Eq']. Only if he can establish this will he be able to infer from the limited information provided about qualia that QR's quale Q has a property which no occurrent property in Eq has. He can then infer that the quale Q does not occur. Hence, QR can be relevantly [and intelligibly] cited as being committed to the claim that if Q is a quale and P is the intelligible PPD item with which Q is identical:

---

18. We can be as specific as this, because QE is at least entitled to restrict the properties in question to those co-occurrent with an episode during which QR claims specifically to be picking out a red quale, or a pain quale, etc., epistemically.

1. X is the property of Q, such that Q is identical with P, and in epistemic situation Eq it is possible to pick out Q determinately, but *not* to determine that Q is a PPD property.[19]

Clearly, QE is not entitled simply to assume that this property X is the diagnostic property of QR's qualia. Consequently, he must find some way of demonstrating that it is, at least to his own satisfaction. The problem is that there is no obvious way available; in fact, it is possible to demonstrate that the position is unsustainable. To see that this is the case, consider firstly the position which QE needs to sustain. QE need only subscribe to the thesis that where the *occurrent* properties mentioned are all those QE discerns during an episode of Eq, in which QR claims to pick out a quale Q:

[A] For any *occurrent* property p which is a PPD property, and the epistemic situation Eq in which p can be picked out determinately, it is possible to determine that p is a PPD property.

And in order to infer that QR's qualia [items of type-[X] do not occur, he can then cite QR's 1 as entailing that [A] is *false* in any case in which 'p' denotes a quale.

It is difficult to see how QE might justify this version of property X. For if X is to be the *diagnostic* property of qualia, QE cannot allow that it is also possessed by any *occurrent* properties discernible in Eq. But if, whenever he identifies his headache in Eq, he also determines that all the occurrent properties discerned in Eq are PPD states, it seems that he already knows that those properties are PPD states. But he could only know *that* if he were able to draw on his prior knowledge that physicalism [at least with regard to the occurrence and contents of Eq] is true. The intelligible difference between QE and QR would then be that QE

---

19. We cannot offer the stronger condition that, for QR, Q can be determinately picked out epistemically even if P cannot, since *ex hypothesi* Q and P are identical.

already knows that [at least in this restricted sense] physicalism is true whereas QR does not. But there must have been a time when QE was able to discern that he had a headache in an epistemic situation Eq, even before he knew that physicalism is true [his attempt to become a physicalist might have produced one]. Hence, his claim that even then his ability to identify his headaches and their properties was always accompanied by his ability to determine that they were all PPD states or properties seems utterly implausible.

Thus, suppose we allow that he has always known that physicalism is true and that in the light of this knowledge he has been able to determine that every p he has identified was a PPD item. We then have to ask him how he discovered that physicalism is true. If he replies that he was able to infer it from his knowledge that every p he identified in Eq was a PPD item his argument is circular. If he replies that he became a physicalist by being in some *other* epistemic state, however, he is thereby conceding that being in state Eq is not a sufficient condition for determining that p is a PPD state. The only other reply available to him, then, is that being in state Eq just is a sufficient condition. But this leaves him back at the beginning, still needing to justify this implausible claim.

The Topic-Neutrality Explanation for Eq.

The situation in virtue of which QR's 1 is true can be explained by reference to the notion of topic-neutral reference. Thus, from the fact that the Morning Star is in fact the object Venus, we cannot infer that if Smith knows that the Morning Star is visible he knows either that the Morning Star is Venus, or even that it is any planet whatever. It is perfectly possible that he should know only that the first of these propositions is true. We can cite a similar case for QE. On the occasion of his headache H he might determine that he has a headache, but not know that H or its properties are any PPD items whatever. The reason why each of these possibilities can occur is that the first referring expression in each case ["Morning Star" and "Headache"] is being taken to be topic-neutral with respect to its referent. Thus, "The Morning Star" is taken to refer to "whatever item presents such-and-such an appearance in a particular epistemic situation" [the details can be filled in as appropriate]. It is only

by construing the expression thus that it remains possible that he does not know, and cannot infer conceptually or logically even in principle, that the Morning Star is in fact the planet Venus.

Similarly for QE and his headache. When, in his ignorance that H is a PPD item, he is able to determine nevertheless that he has a headache H, the referring expression "headache H" must be taken topic-neutrally to refer to whatever PPD item H or properties of H are epistemically available to introspection. It is only by construing the expression thus that it remains possible for him to determine that he has a headache, and that his headache has certain properties, even though he might not know that the items thus discerned are PPD items or properties.[20] An intelligible distinction between the two positions [QE] and [QR] will therefore have to be framed in terms of some other characteristic X which, for QE, no *occurrent* properties in Eq possess.

It is clear from the above discussion that QR and QE can make sense of their respective positions only if they are both taking "discerning in Eq that I have a *headache*", or "discerning that my headache has certain *properties*" to contain a topic-neutral reference to what happen to be, specifically, PPD items. It is this claim which deserves closer scrutiny. In the ensuing discussion, we shall find that there is no remotely plausible way of sustaining QE's position by recourse to his [A]. In order to repudiate QR's position, therefore, QE will need to supplement [A] with some further argument. Specifically, he will need to argue that he has some *additional* information about qualia which enables him to establish that our counterargument is inappropriate. In order to see what sort of information this might be, we can take a more pertinent example.

---

20. At this point we might assume, alternatively, that QE rejects the epistemic possibility envisaged by insisting that he can never determine just that he has a headache. This would render his thesis intelligibly distinct from [QR], but also patently false. He needs some way of making sense of the claim that although there obviously *are* [introspectible] headaches, which are, ex hypothesi, PPD properties, there are no *qualia*.

Thus, suppose that QR undergoes an episode at time t [enters an epistemic situation Eq] in which he claims to be able to pick out a *pain* epistemically, and that he insists that the pain just is the quale Q. But he also claims that he is *unable* to determine in that epistemic situation that Q is a PPD property. This is a permissible assumption in view of the a posteriori nature of the identity thesis to which QR subscribes, and the obvious fact that even QE is able to pick out a *pain* epistemically. QE is then obliged to argue that during the episode in question [or a parallel one in which *he* is able to pick out a pain epistemically], the pain p which he is able to discern in Eq can also be known in that epistemic situation to be a PPD item. Hence, he needs to commit QR to the claim that in the relevant epistemic situation:

> [B] For any *occurrent* pain p which is a PPD property or state, and the epistemic situation Eq in which p can be picked out determinately, it is possible to determine that p is a PPD property or state.

Or, in more eliminative parlance, he could insist that there is no occurrent epistemic situation of the type described by QR. But since [B] is obviously false, it follows that even QE's pain in Eq has the property X in QR's 1. Hence, he is unable to infer from the fact that QR's "pain quale" is supposed to have property X in 1 that X is the *diagnostic* property of QR's pain quale Q.

It seems inevitable, then, that further information about qualia will be needed if QE is to establish that they do not occur. In order to provide that information we might plausibly assume that for QR there is a property of qualia which QE can understand and which enables him to demonstrate that there are no such items. Thus, we might propose that the offending property conceded by QR is that for any occurrent quale Q:

> 2. X is the property of Q, such that Q is identical with P, and in *all possible* epistemic situations in which it is possible to pick out Q determinately it is impossible to determine that Q is a PPD property.

This would evidently provide QE with the information he needs. For now he is able to say that since he is unable to identify any

occurrent property p which satisfies the conditions laid down for Q in 2, he can justifiably infer that qualia are not occurrent properties. Furthermore, it appears that 2 is the minimum condition needed for QE to achieve his objective. For if, contrary to 2, there *were* a possible epistemic situation in which it is impossible to determine that Q and P are identical, the resultant condition would become compatible with the a posteriori identity thesis held by QR. For in that case QR could simply say that the epistemic situation he specified as Eq in 1 is that very situation.

The problem now is that QR would not be committed to 2 either. For QR claims that the quale Q cannot be known *a priori* to be a PPD property; but this allows the possibility that there *is* some conceivable E in which it is possible to determine [a posteriori] that Q is a PPD property. While 2 *might* turn out to be a defensible position, then, it is not the position to which QR is committed. Hence, the description of qualia in 2 cannot be used by QE to distinguish his position from [QR].[21]

The Ontic Commitments of [QE] and [QR].

What QR's thesis boils down to is not that there are any items which QE does not recognise. If the introspected quale Q is identical with the PPD property P, his reference to Q is not a reference to an additional item since, ex hypothesi, Q and P are identical. Nor is it that there is an epistemic route Eq to those items which QE does not recognise. Both have been shown to be committed to acknowledging this route. Thus, both QE and QR must concede that the referring expression "Pain p at time t" refers topic-specifically to a property of the type [Pain], but only topic-neutrally to the type [PPD-property]. We might suppose that each is able to make such a

---

21. QR's a posteriori identity thesis entails just that it is possible to be in an epistemic situation Eq in which it is possible to pick out Q determinately, but not to infer logically or conceptually from the information available in Eq that Q is a PPD property. It is in this sense that he claims the identity relation not to be knowable a priori [in Eq].

topic-neutral reference in virtue of certain neurophysiological functions. We might suppose that their internal information processing faculties are such that they are able to establish that states or properties of a particular type [Pain] are occurring, without also establishing that items of the type [Pain] are items of the type [PPD-item]. Hence, their ability in Eq to determine just that they have a pain. The crucial point now is that the *ontic commitments* of QE and QR with respect to headaches are indistinguishable. Each acknowledges that there are introspectible pains, and that pains are PPD properties; but QR alone claims that headaches are *qualia*. Until some diagnostic property X of qualia can be cited, then, QE is not entitled to say that qualia do not occur, or that epistemic situations of the type Eq described by QR do not occur as described.

The Typic-Classification of Qualia.

QE's thesis can only be intelligibly distinct from [QR] if he claims that there is something wrong with QR's construal of the *typic classification* [Pain]. He has two possible ways of explaining this position. In the first, he must deny that when QR, claims to have identified an item as being of the type [Pain] in Eq he has identified an item as being of any intelligible type at all. In that case, he is eliminating QR's proposed *type* [Pain] because he regards the purported diagnostic property X of qualia [where X is held by QR to be a property of his pain] to be non-occurrent in Eq. But since [or so we are assuming] even QE can determine specifically that he has a pain [rather than a tickle, or a sensation of heat], he is forced to find some intelligible property X which pains do not have. Thus, he must concede that he is indeed able to identify a specific type [Pain], but might insist that items of the type [Pain] do not have property X.

Furthermore, he cannot differentiate his position by pointing out that property X is simply *unintelligible*. For in order to establish that property X is unintelligible he would have to produce some evidence to show that it is not just a property of the type [Pain] which QE finds intelligible. But since we are discussing QE's position in the first-person, this denial lacks any meaning. Thus, if all he is prepared to accept about QR's pain is that it shares

all of its *intelligible* properties QE's pain, it makes no sense for him to say that QR's [Pain] is intelligibly distinct from his own type [Pain]. All he is entitled to say is that there are no intelligible attributes of QR's pain in addition to those which belong to his own pain, and given just this much information the two positions are indistinguishable. QE might insist that *if* QR's pain is taken to be of some unintelligible type distinct from his own type [Pain] it does not occur; but this is a claim to which QR will happily subscribe. QR claims that his own type [Pain] is intelligible. Thus, QE's attempt to distinguish his position from QR's along these lines must be a failure.

Finally, then, let us assume that QE *does* understand the further stipulation that QR's qualia [of which his pain is an instance] are of some *qualitative* type-[X]. In fact, let us assume that he understands quite a few of the claims which QR makes about his supposed qualia. He understands all the information about qualia there is to understand, with the one exception that he is unable to pick them out epistemically in his own case. Thus, we might offer an expanded account of the property X which qualia are supposed to have. Firstly, as before, and taking Q to be a pain quale:

1. X is the property of Q, such that Q is identical with P, and in epistemic situation Eq it is possible to pick out Q determinately, but *not* to determine that Q is a PPD property.

And in addition, something like:

3. Qualia are the identifiable and distinguishable qualitative properties of experience which can occur and be epistemically picked out [in introspection] irrespective of whether the subject has any sensory input from the external world. Also, they are subjective, in the sense that there is a type of epistemic situation [Eq] in which it is possible to pick out one's own qualia, but no epistemic situation in which it is possible to pick out the qualia experienced by others.

Not all qualia-reductivists might agree with this account, but this is beside the point. The point is that apart from the qualitative properties of experience *per se*, all of the facts which QE is

required to understand in 1 and 3 are likely to be relational or epistemic properties of the sort he *would* understand when applied to PPD items, even if no PPD items have them. Hence, any account of qualia which employs properties of this sort should at least be understood to this extent. Suppose, then, that there is a qualia-reductivist who does agree at least broadly with this account. Thus, according to 3, he might maintain that he experiences qualia when he is either dreaming or hallucinating, and that he is able to pick out and distinguish the particular qualia he experiences. He claims that he is able to determine that he experiences a particular quale Rp, and that it is the sort of quale he typically experiences when looking at ripe tomatoes. Similarly, he is able to determine that he experiences a particular quale Gp, and that it is the sort of quale he typically experiences when looking at a thriving lawn. He might even agree that while it is possible to pick out one's own qualia determinately in introspection it is impossible to pick out the qualia experienced by others determinately by any means whatever. Furthermore, in accordance with 1, he claims that although they are identical with some particular PPD items, epistemically it is possible for him to recognise and distinguish his own qualia without also being able to determine that they are identical with any PPD items whatever.

The point is just that QE can still understand at least some of the relational and epistemic properties attributed to them in 1 and 3. If he then construes qualia *topic-neutrally*, as whatever items [qualitative properties of experience] are supposed to have all of the *intelligible* properties described in 1 and 3, there is nothing in the above account which he will fail to understand, other than the topic-neutrally specified qualia themselves. It seems, therefore that he might then reject QR's "qualia" as being properties of a type which simply does not occur.

Thus, when QR says that he is experiencing a quale, his report must be construed as the claim that he is in an epistemic situation Eq with respect to P. But since we have established that even for QE there is such an epistemic situation [in the case of headaches, for example], he can only intelligibly distinguish his position from [QR] if he can cite a *further* intelligible characteristic X of QR's qualia as picked out determinately in Eq. He can then hope to distinguish his own position intelligibly from [QR] by claiming that

the properties picked out determinately in his own occurrent Eq as headaches, for example, lack X. Taking this line, then, his rejection of qualia will turn out to amount, not to a discrepancy in ontic commitment, but rather to a disagreement with QR over the *properties to be ascribed to qualia*. As we have already explained, this might be legitimately regarded as a form of eliminativism. For it is not at all clear that there is an intelligible distinction to be drawn between claiming that the properties referred to as "qualia" *do not occur*, and claiming rather that the properties referred to as qualia *are not as described*.[22]

Although QE must find his own occurrent type [headache] intelligible, then, he now has two possible ways of trying to distinguish his position from QR's. Construing H as the distinguishing property of his own headache, discernible even by QE in Eq as a headache, he can claim either that:

> [C] X cannot be *intelligibly* ascribed to H [even though X is intelligible].

or that:

> [D] X cannot be ascribed *in fact* to H.

He could only support [C] if he could cite some characteristic X which, for QR, the headache discernible in Eq must have, and which QE at least finds intelligible. If QR is committed to his H having X, QE could then try to establish that such an ascription to his own occurrent headache is itself unintelligible. His eliminativist position would then amount to the claim that there are occurrent epistemic situations of the type Eq, but that it is *unintelligible* to propose that the headache discernible in Eq has characteristic X. To take an absurd example, suppose that QE has identified X as being the number seven. He could then argue that QR's commitment to the claim that a quale is the number seven is simply unintelligible when applied to his own headache. But in order to substantiate that claim QE would then have to cite some intelligible characteristic Z of his

---

22. See chapter II, and footnote 16 of the present chapter.

occurrent headache which renders the claim that his occurrent headache is the number seven *unintelligible.*

Ultimately, then, QE's charge of unintelligibility must be grounded on the citing of an *intelligible* characteristic Z of his occurrent headache. He must show in accordance with option [C] that X cannot be intelligibly ascribed to his headache because his headache has property Z. He could say, for example, that X is the property of being the number seven and Z is the characteristic of being an introspectible property. He could then point out that the ascription of the number seven [which has X] to his occurrent headache [which has Z] would be unintelligible. It would be unintelligible to claim that an introspectible property is the number seven. He would then have established that [C]; some intelligible characteristic [X] of QR's qualia cannot be intelligibly ascribed to his own occurrent properties. Effectively, he will have demarcated his position intelligibly from [QR] by claiming that the occurrent properties which he is able to discern in introspection [in Eq] do not have X; or, in more eliminative style, that there is no occurrent property of the type [X] in Eq.

If QE is to adopt position [C], he is therefore obliged to show that some characteristic Z of his occurrent properties in Eq ensures that QR's intelligible characteristic X cannot be intelligibly ascribed to those properties. Clearly, the property X already described in 1 will not be suitable for his purposes, since QE and QR [must] agree that they are able to pick out an occurrent headache introspectively without also being able to determine that it is a PPD item. In that respect, therefore, the two theorists are in agreement.

We can quickly see, however, that some of the most likely candidates for X will fail to afford QE the discrepancy he seeks. Thus, QR *might* be prepared to claim that his Eq is the state in which, in addition to affording introspective access to PPD items of type [P] under the epistemic conditions set out in 1, at least some of the properties described in 3 can also be ascribed to his qualia. Thus:

(i). Occurrent qualia are *experienced.*

(ii). A *qualitative character* is experienced.

And, more generally,

(iii). Qualia are epistemically available intrapersonally, but not interpersonally [i.e, they are essentially discernible only subjectively].

The claim in (i) exudes a prima facie air of interest, since it is not at all obvious how a physicalist might explain how experience can be accounted for within the confines of his philosophical position. By the same token both (ii) and (iii) would seem to present at least a considerable challenge for him. For if physicalism is construed in some coherent terms as the thesis that the objective facts, or the facts afforded by "the view from nowhere" [Nagel, 1986], are the only facts there are, all three of the above ascriptions would appear to present a formidable challenge to that thesis. So it appears that we now have at least a prima facie basis for drawing an intelligible distinction between [QE] and [QR]. [QR] might incorporate any or all of the propositions (i) - (iii), while [QE] simply rejects them as being false.

There is good reason to suppose, then, that any *plausible* version of [QR] will be unable to incorporate proposition (iii). Thus, we assumed at the beginning of the chapter that physicalism *per se* is to be construed as incorporating the thesis that every occurrent property just is a PPD property, and that all PPD properties are both conceptually and epistemically available in the third-person perspective. If that is the brand of physicalism we choose to adopt, then, it is logically impossible that (iii) should be true. QR cannot say that his qualia conform to this concept of physicalism and yet qualia are epistemically private. For in that case he would be committed to the occurrence of an epistemically available property of qualia [epistemic privacy] which would render qualia non-physical. So while it is *intelligible* to suppose for QR (iii) is true, it would be logically impossible for him to do so within the confines of the brand of physicalism we have chosen.[23]

---

23. This leaves open the possibility that some other brand of physicalism might be adopted, in accordance with which QR might consistently subscribe to (iii), and a distinction between [QE] and

The question of whether QR might subscribe to propositions (i) and (ii) is less straightforward. Together, they amount to the claim that a quale Q is the qualitative character of a PPD item P discerned in experience. If Q is discernible in experience, then, our physicalistic strictures dictate *both* Q and an episode of experience [being in Eq] should be epistemically and conceptually available within the third-person perspective. In this case, however, there is room for interpretation of QR's position. Thus, if he intends "the qualitative character of an experience" to refer to a property which our physicalism precludes, his position will be logically untenable. If, however, he intends "an experience of qualitative character" to refer to an episode which conforms to our physicalistic constraints, we can take him to be saying, uncontroversially, that such an episode is just a PPD episode. Taking the logically permissible interpretation to be the one intended, then, QR can only be said to differ from QE if QE at least rejects, as non-occurrent, an epistemic episode of the sort which QR describes as the episode in which he discerns the quale which he describes as being both a PPD property and "the qualitative character of an experience". In order to determine whether there is any intelligible distinction to be drawn between [QE] and [QR] in this respect, then, it will again be necessary to determine *which* [intentionally inexistent] property is being referred to as a quale.[24]

This presents us with an apparently insoluble problem. Consider, for example, Smith's headache H at time t. Since, ex hypothesi, QE and QR are in disagreement as to which PPD properties occur during this episode, if QR claims that the headache Smith experiences in introspection [in Eq] just is an occurrent PPD property P, the only

---

[QR] drawn on that basis. For now, however, we are attempting to draw that distinct within the confines of the version of physicalism already explained.

24. The only distinguishing feature of QR's "experience" Eq so far available to QE is that Eq is the epistemic situation in which qualia are discerned. Hence, QE can only eliminate Eq if he can show that the occurrent epistemic situation is not of this type.

intelligible rejoinder for QE is that it is not. As we have seen, such a denial can be reconstrued without loss of import as the claim that there is something wrong with QR's characterisation of P; that he ascribes properties to P which no occurrent property discernible in Eq has. At the same time, however, we have argued that QE must at least acknowledge that an occurrent PPD property of the type-[headache] *is* discernible in Eq. So clearly QE cannot be disagreeing with QR in that respect. Instead, he must be claiming that QR has simply misdescribed the character of P. He must insist that while Smith's headache is a PPD property P, it does not have some characteristic X in virtue of which it would count as a quale. And this entails that QE and QR are *not*, as was originally supposed, in agreement as to the contents of [S]. QR claims that a property with characteristic X is occurrent, while QE claims that it is not.

What we find, then, is that the proposed distinction between [QE] and [QR] is logically incompatible with the original set of assumptions. On the understanding that the properties referred to are all [intentionally inexistent] *physical* properties, the suitably modified assumptions are that:

1. The occurrent properties are to be referred to as the set of PPD properties [S].

2. Logically prior to the debate about qualia, QE and QR agree on which properties belong to [S].

We have just seen that if QR is to distinguish his thesis intelligibly from [QE] he is committed to conceding that the agreed contents of [S] are not exhaustive; that there is an occurrent property X of Smith's headache which is not a member of the agreed set [S]. Hence, if QR accepts 1, he must reject 2. In other words, QR claims that there is a PPD property Q [a property defined as having a particular PPD characteristic X] and QE claims that there is no occurrent PPD property which has that characteristic. But this, again, leaves open the question of *which* characteristic X is in dispute.

In terms of our specific example; if QE and QR agree that in introspection Smith is able to determine that he has a headache [an instance at time t of the PPD property H], then the property in

dispute must be some *other* property X which QR claims to be possessed by H but QE does not. If QR claims that H just is the quale Q, it follows that the dispute is not over the occurrence of property Q after all. It is, rather, a dispute over how Q is to be described; which properties should be assigned to it. Hence, QE is a *reductive* physicalist with regard to Q, but an *eliminativist* with regard to property X. Or, he is an eliminativist with regard to the Q which has X. Unless some property X can be specified, then, such that QR claims it to be a PPD property of qualia and QE claims that it is not, there is no intelligible distinction to be drawn between the two positions. Suppose, to the contrary, that QR does *not* hold H to be the quale Q, but holds instead that H has the property Q as a distinguishing characteristic. The same problem still remains. For now QE can be said to deny the occurrence of an H which has Q, and the dispute only has any meaning if Q can be intelligibly specified. In either case, then, the dispute is over the appropriate description of Smith's introspected headache H. In eliminativistic terminology, it is over the question of whether [S] contains a property as described by QR.

The only way of resolving the contradiction is by rejecting the original assumption that there is agreement over the membership of [S]. QR claims that there are PPD properties of type-[H] which have characteristic X, while QE denies this. He says that there are instances of H, but that they do not have X.

Summary.

In summary, then, the position is as follows. QE's best chance of distinguishing his position from [QR] seemed to consist in showing that QR claims that a particular type of epistemic situation occurs, whereas QE claims that it does not. Hence, QE will insist that the properties supposedly available in that epistemic situation do not even occur. Thus far, the proposed distinction appears to be unequivocally ontological in character. Once we see that introspective access to pains, for example, is undeniable, the dispute becomes that of *which properties are available in that situation*; how introspectible pains are to be described. The ontological character of the dispute is then less clear. Instead, it becomes more plausible to regard it as a disagreement over the

character of the [undeniably] introspectible properties. But we have yet to find any characteristic X in terms of which [QR] can be determinately distinguished from [QE]. We saw that there *are* versions of [QE] which would be intelligibly distinct from [QR], but they would be utterly implausible. Thus, if QE claims that there is no epistemic situation Ep in which Smith can discern that he has a headache, but not that it is a PPD property, his position would be intelligible, but clearly false. Similarly, there are interpretations of [QR] which would provide the required distinction. Thus, if QR were to claim that qualia are epistemically private, or have a qualitative character which is discernible only in the first-person perspective, his position would be intelligibly distinct from [QE]. But at the same time his resultant position would conflict with our common-sense version of physicalism, as an essentially third-person account of the world.

Dennett's Account of "Seeming Phenomenology".

One further way in which QE might try to differentiate his position from [QR] within the constraints of our common-sense physicalism is as follows. Thus, he might insist that some sort of intelligible *illusion* occurs even for him, but that whereas QR succumbs to that illusion QE does not.

Dennett, for example, claims that although phenomenal properties, or qualia, do not exist per se, they do at least *seem* to exist. The following passage gives a clue as to how Dennett proposes to explain the illusion.

> There seem to be qualia, because it really does seem as if science has shown us that the colors can't be out there, and hence must be in here. Moreover, it seems that what is in here can't just be the judgements we make when things seem colored to us. This reasoning is confused, however. What science has actually shown us is just that the light-reflecting properties of objects cause creatures to go into various discriminative states, scattered about in their brains, and underlying a host of innate dispositions and learned habits of varying complexity. [Dennett p372].

In terms of Smith's headache discerned in Eq, the weakest interpretation of any interest, then, would be that when Dennett

concedes that we at least seem to experience, say, the quale Q172, the following state of affairs obtains. Firstly, we do not experience Q172, since it does not occur. But it *seems* to occur. That is, a physical state T172 occurs in which a subject is at least inclined or tempted to *believe* that he is experiencing Q172.

Suppose, then, that seeming to experience Q172 is nothing more than *believing* that Q172 is discerned in Ep. Without having to get involved in the contentious issue of whether a state of belief *per se* can be fully accounted for in PPD terms, it is clear that if the belief is intelligible then so too must the object of the belief, Q172, be intelligible. If it is intelligible, there would seem to be some prospect of distinguishing intelligibly between [QE] and [QR] in the following way. QR holds the belief that Q172 occurs, while QE does not. But this proposal leads to an immediate problem. For if Dennett's illusion amounts to no more than holding the belief that Q172 occurs, it is not the sort of illusion in terms of which the property Q172 might itself be rendered intelligible. Consider, for example, Smith's belief that God exists. If Dennett encourages Smith to see that the existence of God is just an illusion, and that the illusion consists just in believing that God exists, the question remains as to which type of [intentionally inexistent] item "God" refers to. Similarly, then, if seeming to experience Q172 consists just in believing that Q172 occurs, we are no nearer to an understanding of which type of intentionally inexistent property "Q172" refers to. In essence, we are back at the beginning, still needing to find some intelligible characteristic of occurrent headaches which QR subscribes to but QE does not.

Suppose now, then, that Dennett were to provide further information about the *content* of the belief. Seeming to experience Q172 amounts to believing that Q172 occurs, and in addition, Q172 can be specified intelligibly as some property in addition to "the judgements we make when things seem colored to us" [in Dennett's example]. Now we have seen that the positing of such a property by QD would render his thesis intelligible from physicalism *per se* in the following way. QD claims that the physicalistic account of experience is complete while QD denies that this is the case. In order to *substantiate* his intelligible position, then, he will have to cite an intelligible non-physical property which does occur. The distinction between [QE] and [QR], however, is not so readily drawn.

Suppose that there is some intelligible sense in which qualia can be *believed* by QR to occur in introspection. In that case, we might reasonably assume that QR believes Q172 to occur in both QE's and QR's introspection. The difference between the two theorists is then that while QR holds that belief, QE does not. QR holds that the property Q172 is an occurrent PPD property, while QE holds that it does not occur. Hence, in order to make any sense of QR's belief, some independent specification of the Q172 which QR believes to occur must be provided. Casting the theoretical disagreement between QE and QR as a difference of belief therefore brings us no nearer to understanding the object of that belief; QR' quale. The belief that Q172 occurs can still only be intelligible if Q172 can itself be specified intelligibly. And this leaves us back at the beginning, still trying to understand the diagnostic characteristic X of qualia.

Clearly, then, the appeal to a belief *per se* gets us no nearer to an understanding of the sort of property which QR believes is possessed by headaches. An intelligible account of Q172 per se, as the object of the belief, must still be provided. Until such an account can be provided, we have no common currency in terms of which to differentiate the respective claims of QE and QR.

Conclusion.

We established that there is at least an intelligible difference between the predicates applied by each theorist respectively to phenomenal properties, or qualia. Thus, whereas for QR qualia *occur*, for QE they *do not occur*. But the two positions boil down to a shared commitment to the occurrence of an epistemic state Eq in which certain PPD states or items can be determinately identified [as headaches], but not as PPD items of any kind. The hallmark of that epistemic state is the ability of the subject in that state to identify PPD items as being items of a particular type [headache], even though that type is not known to correspond to any PPD type whatever. The only intelligible distinction between the two positions must therefore be drawn in terms of what the rival theories say about the headaches discerned in Eq. For QR it must be an item of some particular PPD type, while for QE it must be an item of no particular PPD type, and therefore, since for the physicalist

PPD types must be the only occurrent types, an item of no particular occurrent type at all. But since even QE is able to determine whether he has a headache or indigestion, for example, he is evidently able to distinguish between the two types. Thus, although QE claims that he alone is resisting Dennett's inclination to posit the occurrence of qualia in such a situation, we have yet to understand what sort of properties qualia are supposed to be. And at this stage, this leaves [QE] indistinguishable from [QR].

The necessity for this clarification is underlined by the vagueness exhibited by both Dennett and Rorty with regard to qualia-discourse. Thus, for Dennett, Otto's reference to "my quale" is on occasions construed as an unwitting reference to a [neurally grounded] complex of dispositions. He explains, for example, that:

> What qualia are, Otto, are just those complexes of dispositions. When you say "This is my quale," what you are singling out, or referring to, whether you realise it or not, is your idiosyncratic complex of dispositions. You seem to be referring to a private, ineffable something-or-other in your mind's eye, a private shade of homogeneous pink, but this is just how it seems to you, not how it is. [Dennett 1991 p389].

Rorty takes a parallel view when he asks rhetorically, "..cannot we see that our talk of mental states was [note the past-tense] merely a place-holder for talk of neurons?" [1979, Rosenthal p272]. In footnote 24, p 286, however, he acknowledges at least that there is less of a difference between eliminative and reductive materialism than he had previously thought [in Rorty, 1970].

We can see immediately that the position declared by each writer here is non-committal with regard to the distinction between qualia reductivism and qualia eliminativism. For while each subscribes explicitly to the physicalistic thesis that there are no properties other than physical or dispositional properties, there are shades of both eliminativism and reductivism in each account. Thus, while Dennett claims, in reductive fashion, that a reference to qualia is in fact a reference to dispositional traits, he also points out in eliminative fashion that the qualia referred to do not have some of the properties ascribed to them; that the qualia bearing those properties do not occur. Similarly, Rorty construes qualia discourse

here as referring to mental states, but in the footnote cited is evidently unsure as to whether it refers to anything at all.

Dennett asks "Are qualia functionally definable?" His answer is "No, because there are no such properties as qualia", "Or, yes, because if you really understood everything about the functioning of the nervous system, you'd understand everything about the properties people are actually talking about when they claim to be talking about their qualia" [1991, pp 459-460]. The crucial point is that although logically incompatible predicates can be ascribed to the respective theses, they are both held true by Dennett and Rorty only on the assumption of an *equivocation* over the referent of "qualia" in the two cases. Thus, if "qualia" refers to properties other than PPD properties, they do not occur, while if it is taken to refer to PPD properties they do occur. Thus, the eliminative and reductive predicates are held to be compatible just because they are being applied to *different properties*. In this chapter we have argued that unless some identifying characteristic of qualia can be specified determinately, there is no further sense to be made of a [QE]/[QR] distinction which does not depend on such equivocation. The two will be logically complementary theses subscribed to by our common-sense physicalist. Quine summarises this position in the following way.

> Corresponding to every mental state .... the dualist is bound to admit the existence of a bodily state that obtains when and only when the mental one obtains. The bodily state is trivially specifiable in the dualist's own terms, simply as the state of accompanying a mind that is in that mental state. Instead of ascribing the one state to the mind, then, we may equivalently ascribe the other to the body. The mind goes by the board, and will not be missed. [Quine 1985, Rosenthal p287].[25]

Why, then, do we consider it important to draw an intelligible distinction between the two theses, when our principal interest is in evaluating the credentials of *reductive* physicalism?

---

25. The *physicalist* Quine might again be understood as implying here just that "qualia", for example, occur if they are construed as bodily states or properties which do occur during episodes of

Put simply, the reason is this. Reductive physicalism holds to the thesis that qualia are occurrent physico-dispositional properties. Qualia *dualism*, on the other hand, claims that qualia are occurrent *non*-physico-dispositional properties. Hence the two positions are explicitly contradictory. Now if the contradiction is to amount to any intelligible disagreement at all, we must assume that the two theses are referring to the *same qualia*. But if we cannot distinguish intelligibly between the [reductive] claim that qualia are *occurrent* properties and the [eliminative] claim that they do *not* occur, it is clear that we have no idea which qualia are being referred to. Hence, the distinction between [QR] and [QD] will be unintelligible too, except insofar as QD claims that the physical account of experience is incomplete. Unless the domestic dispute among physicalists is intelligible, then *a fortiori* there is no intelligible distinction to be drawn between the thesis that qualia are occurrent PPD properties and the thesis that they are occurrent non-PPD properties.

Nevertheless, it remains possible that QE and QR *should* be able to determine which particular properties they are referring to when they disagree over the occurrence of qualia. If they can do so, the compatibility of the two theses as subscribed to by Dennett, Rorty and Quine will no longer obtain. For in that case it will make sense to treat physicalism and qualia-dualism as mutually incompatible theses about a particular property. As we have made no progress towards a specification of the qualia in question, then, and yet the possibility of doing so has not been ruled out, we might look elsewhere for the required information. For this, we turn our attention to the thesis of the qualia-dualist. Since he claims [intelligibly] that we experience *irreducibly non-PPD* qualia, his thesis can only be *justifiable* insofar as the qualia he refers to can be specified intelligibly. In subsequent chapters we shall be exploring some of the principal arguments he might offer in support of his thesis. In so doing, however, we do not presuppose that qualia have been rendered intelligible. Rather it will become

---

sensory input, or at other times when hallucination or illusion produces those states or properties, while they do not occur if they are supposed to be anything else.

apparent that the very *specification*, or diagnostic properties, of the qualia referred to by the dualist might, at least to some extent, be inferred from the methods he employs in an attempt to show that they are irreducibly non-physical.

In the following discussion, then, we shall acknowledge that there is as yet no intelligible distinction in the first-person between the positions [QE] and QR] by referring to them jointly as "qualia-physicalism", [QP]; the thesis which entails that even when we are able to identify a quale in the first-person, we might nevertheless be in an epistemic situation Eq in which it is not possible to determine that the item identified is a PPD item, even though it is. Whether the items thus identified are occurrent *qualia*, in the sense intended by the qualia dualist, is then a further question to explore. Equivalently, but in a less ontological idiom, the question will be whether the items we do discern in introspection are correctly described as being physical or non-physical properties. The thesis that the qualia we are able to identify are irreducibly non-PPD items, and therefore that there is a type of occurrent item distinct from the physical type, will be referred to as "qualia-dualism", or [QD].

Chapter IV

THE INVERTED SPECTRUM ARGUMENT

Introduction

The inverted spectrum argument is to be considered here as one possible argument for the occurrence of irreducibly non-physical properties, thus supporting the qualia-dualist's thesis that the physicalist's account of sensory experience is incomplete. If it proves to be intelligible, we might legitimately infer that, despite our previous failure to distinguish intelligibly between [QE] and [QR], it really does make sense at least to suggest, with QD, that qualia are the non-physical properties uncovered by the inverted spectrum argument. It will then be an *intelligible* matter for further investigation as to whether the qualia thus identified might be occurrent physical or non-physical properties, or not even be occurrent properties at all.

As Dennett observes, intuitions which suggest an inverted spectrum possibility can be found in the work of John Locke. Thus, Dennett reflects in Lockean fashion that:

> There are the ways things look to me, and sound to me, and smell to me, and so forth. That much is obvious. I wonder, though, if the ways things appear to me are the same as the ways things appear to other people. [Dennett, 1991, p 389]

In its conventional form, the inverted spectrum argument is intended to establish that there is a conceptual distinction to be drawn between qualia and PPD items, and hence at least a distinction in terms of which [QD] can be rendered intelligible. But we must be careful to consider at the outset what the purported distinction does and does not entail. If we concede that Locke's speculation is at least intelligible, we might permit the further assumption that we can also imagine our *own* experiences [i.e., intrapersonally] having qualitative characteristics other than the ones they do have. If so, we should then be able to infer at least that appearing [qualitatively] the way red objects appear *to me* [call this property Rp] is conceptually distinct from *appearing the way red objects appear per se*. It would be possible to imagine red objects looking

the way green objects, for example, actually look to me. And this possibility would confound the claim that perceiving that an object looks red involves nothing more than discerning sensorily that the object looks *like a red object*. For there must, in addition, be *a way* that red objects look to me, which can be imagined to have been other than it is. Thus, the conceptual distinction might be of the sort outlined in chapter I between the *qualitative* content of appearance and the intentional or *representational* content of appearance.

But the question we are now interested in is quite different. Thus, is the way red objects look to me *qualitatively*, [Rp], something distinct from the physical state or property [RP] induced in me when I see a red object? Regarding our powers of imagination, we can imagine *red objects* looking in some way other than Rp, but can we imagine Rp not being *RP*? The inverted spectrum argument suggests that we can; that we can imagine an occurrence of RP co-obtaining with some other property rather than Rp, and hence Rp not being RP. For example, we can imagine the way red objects appear as being other than it in fact is [Rp]; that it is the way green objects actually appear [Gp], while the *occurrent RP* remains invariantly induced in us whenever we look at red objects. If we can do that, we might infer that the way each colour appears qualitatively will be *conceptually* distinct from the physico-dispositional states and properties induced in us when seeing the appropriate colour.

In the context of the present discussion, it would be tempting to suppose that this state of affairs has already been shown to obtain. Since there is a possible epistemic situation Eq in which we are able to determine that we are experiencing quale Rp even though we *cannot determine* that Rp is any PPD item whatever, it already seems to follow that Rp and all PPD items are conceptually distinct. Thus, Eq can only obtain in virtue of the fact that the type [Rp] is only topic-neutrally related to the type [PPD]; that the referring expression "Rp" refers topic-specifically to items of the type [Rp], but only topic neutrally if at all to items of any PPD type [RP]. Hence, the [physico-dispositionally] topic-neutral concept of a property Rp, as discerned in introspection, must be logically independent of the topic-specific concept of RP. An occurrence of Rp does not logically or conceptually entail an occurrence of RP, and hence Rp must be logically and conceptually distinct from RP.

While this observation was sufficient to draw attention to a common thesis of [QE] and [QR], that the proposed identity thesis can be substantiated only *a posteriori*, however, it is *not* sufficient to secure a *conceptual* distinction between qualia and PPD properties of the kind required by the inverted spectrum argument. What we are now interested in finding out is whether Rp and RP are *in fact* distinct properties, and this is precisely what the occurrence of Eq fails to establish.

If the objective of the inverted spectrum argument is to show that Rp and RP are in fact distinct properties, then, it will have to be at least *intelligible* to suggest that Rp is distinct from RP in fact. Clearly, it is intelligible. For it makes plain sense to suggest that the property Rp which we discern in introspection is not a physical property at all, just as we saw earlier that it makes sense to suggest that in Eq we can discern an occurrent headache, as a property of the specific type [Hp], without even knowing that it is a physical property. In that sense, then, the two phenomena may be said to be conceptually distinct. But in order to infer from our imaginative powers that the two phenomena are *in fact* distinct, a conceptual distinction of a stronger variety is required. For, consider; if two referring expressions refer to a common referent it will not be possible to imagine the common referent [whatever it happens to be] being other than itself. Only if they are distinct, therefore, might we reasonably suppose that it will be possible to imagine an occurrence of one which is not an occurrence of the other. So in order to show that they are distinct, we need some way of establishing *at least* that it is possible to imagine, topic-specifically, an occurrence of Rp which is not an occurrence of RP. The occurrence of Eq shows just that:

1. We can imagine that Rp [as picked out topic-neutrally in introspection] should turn out not to be any PPD property RP.

But if imaginability is to be the test, we *need* to establish that:

2. We can imagine [topic-specifically] an occurrence of Rp which is not an occurrence of RP.

So it is important to distinguish between the *epistemological* state of affairs imagined in 1 and the *metaphysical* state of affairs

imagined in 2. As we saw in chapter III, 1 can be true simply in virtue of our ignorance or limited grasp of physical theory in Eq, but 2 can only be true if Rp is not a PPD property; the very fact in question.

In order to establish 2 we might employ the inverted spectrum argument; firstly to establish that Rp and any PPD property RP can indeed be imagined not to co-occur, thus establishing 2, and thence to infer that the property Rp discerned in introspection is in fact distinct from any physical property RP.

Initially, then, the aim of the inverted spectrum argument must be to establish that a property Rp can be imagined *topic-specifically* not to co-occur with any specific PPD property. If we can do that, it will follow that we can imagine an occurrent Rp which is not the particular PPD property with which the reductive physicalist holds it to be identical. Thus, premise 2 will have been verified.

The standard argument sets out to establish that, for any quale Rp (property of type [Rp]) and any PPD property RP (property of type [RP]) with which Rp is supposed to be identical, and some other quale Gp and some other PPD property GP with which Gp is supposed to be identical, at least one of the following conditions obtains:

[1]. RP can be imagined to co-obtain with either Rp or Gp.

[2]. GP can be imagined to co-obtain with either Rp or Gp.

[3] Rp can be imagined to co-obtain with either RP or GP.

[4] Gp can be imagined to co-obtain with either RP or GP.

Since Rp and Gp are of distinct types, and similarly RP and GP are of distinct types, each of these claims implies on its own that neither Rp nor Gp can be of the same type as its proposed PPD correlate. Hence, the conceptual version of the inverted spectrum argument need only show that any one of them is true in order to achieve its objective. If Rp and RP can be imagined not to co-obtain invariantly, for example, the topic-specific concept of an item of type [Rp] cannot be the concept of an item of type [RP]. As we saw in the introduction that our version of physicalism entails that all

introspectible types must be PPD types, then, it will follow that our reductive physicalism is false.

Notice that in response to the proposed PPD analysis of experiential qualities we are initially discussing conceptual identity in particular here. One problem with trying to find a good example of a conceptual identity, however, is that we have not actually defined conceptual identity in the first place. Indeed, if the vast quantities of literature devoted to the clarification of the concept of analytic truth are any indication, we are still a long way from having such a definition available to us [see Quine, *Two Dogmas of Empiricism*]. Clearly, it would be inappropriate to address this problem in full in the context of an evaluation of the inverted spectrum hypothesis. For present purposes, rather, we will be assuming simply that "conceptual identity", as required for the purpose of determining the content of Smith's concept of qualia, is roughly the sort of identity between topic-specific properties p and P which cannot be imagined to fail. Although this assumption might require further refinement, it at least lends structure to our discussion and enables us to follow the inverted spectrum argument as expounded by its proponents.

The Basic Argument.

As our paradigm, we shall consider the intrapersonal version of the argument [*adapted* from Block, *Inverted Earth*]. In the argument we are about to consider, the proposal under consideration is that qualia are to be characterised in terms of overtly dispositional or functional traits. Naturally, for the physicalist, these traits will be neurophysiologically grounded, but the claim under consideration is that the specific *properties* [construed as universals] referred to as qualia can each be dispositionally characterised. The suggestion that qualia might be characterised in less overt, neurophysiological terms, will be explored later.

Smith has a lens transplant, after which the qualitative content of red and green experience, respectively, is inverted for him. He now experiences red things qualitatively as he used to experience green things, and vice versa. Nevertheless, the dispositions originally associated with his seeing objective samples of red either remain

constant, or are eventually restored. The four stages of the experiment are;

> Stage 1. Smith is dispositionally and qualitatively normal.
>
> Stage 2. Lenses installed. Qualitative inversion occurs. Period of dispositional wavering [for example, he knows that grass is green but experiences it as looking the way red things used to look - he therefore has difficulty deciding whether to use "red" to refer to the property of the object producing that qualitative content].
>
> Stage 3. Smith's dispositional state reverts to normal. Falls in with social usage again [apart from the fact that he is wearing the lenses], saying "that is red", and even "that looks red" when looking at a red object. But still remembers the previous qualitative content of seeing red.
>
> Stage 4. Forgets or ignores the qualitative content of seeing red at stage 1. Back to normal dispositional state, but with inverted spectrum.

The conceptual implication of this scenario is supposed to be that it is possible to imagine that Smith undergoes an inversion of his qualia; that he experiences Gp rather than Rp, even though his dispositional state reverts to the state RP normally associated with experiencing Rp. Hence, since this implies that [1] is imaginable, it entails that qualia Rp and Gp must be conceptually distinct from any dispositional state whatever. If this version of the inverted spectrum argument is sound, the concept of Smith experiencing a particular quale Rp or Gp cannot be expressed in overtly dispositional terms. Clearly, a parallel conclusion can be drawn with respect to any other quale Smith might experience.

Before going into the credentials of the scenario cited by the inverted spectrum argument, then, we can already see in broad terms how the qualia in question are to be characterised in support of [QD]. They are to be the "experiential character, or "what is like, experientially" [see Nagel, Jackson, Robinson, in chapter VI], of seeing colours or having other sensory perceptions. Thus, Rp is to be the experiential character typically associated with seeing red, while Gp is to be the experiential character typically associated with seeing green. Similarly, we might say that *phenomenal pain* is the experiential character of having a pain. It should be noticed

that according to this account, qualia-dualism stands to be corroborated only if the qualia thus cited are specifiable *topic-specifically*. Thus, if "what it is like" is construed topic-neutrally just as "whatever property Rp discerned in introspection happens to be", the fact that the property in question can be imagined, in Eq, not to co-obtain with RP does not entail that the two are in fact distinct. What the inverted spectrum argument needs to establish is that there is an introspectible property Rp which can be *topic-specifically* imagined to occur without co-obtaining with any PPD property RP. [26]

---

26. Some commentators tend to regard the inverted spectrum argument as purporting to provide *substantiation* for the claim that introspectible properties occur. Christopher Hill, for example, takes it to be that after spectral inversion has occurred:

> ...objects no longer have the same effect on [Smith] as they used to have. He will be able to tell from the inside that a spectrum inversion has occurred. But this means that he will be aware of the intrinsic natures of his sensations. [Hill, p197]

His response is that:

> [Smith] can determine that a spectrum inversion has occurred without taking note of his sensations. He need only take note of the appearances of things. For example, he can determine that an inversion has occurred by observing that the things that used to look blue now look yellow. [p 197]

The crucial question, of course, is *how* he would be able to observe that the colours of objects appear to have changed. The inverted spectrum argument *presupposes* that introspectible properties of particular types, [e.,g. Headache, Pain, What it is like to see red, etc.], do occur, in order to explore the nature of those types. In an attempt to avoid acknowledging that "the intrinsic natures of his sensations" [or in our account, his ability to recognise which quale is the object of his belief] enable him to do so, Hill explains that the relevant information is determined *inferentially*, via collateral considerations. Thus, to vary the example, he is prepared to insist that if he is at the top of a tall building viewing the people far

below, in order to determine that his visual image of one of those people occupies a small portion of his visual field he must first endure the following mental gymnastics:

> Having observed, for example, that the people on the ground look like ants, one may go on to affirm that the sensations by which the people are represented occupy comparatively small areas of the visual field. As I see it, this proposition about the visual field can only be obtained by inference. It would be a mistake to think that it is shown to be true by the data of immediate awareness. [Hill, p 199]

He prefers to eliminate reference to such data *per se*, saying rather that we really do determine that we have a particular belief about the character of our experiences by way of a complex inferential route of the sort just described. Thus, for Hill, the belief that one is experiencing a small visual image is to be inferred from his initial recognition of the fact that people look, in the relevant respect, the way ants look when viewed more closely. Having drawn on his memory of ant-viewing episodes he then proceeds to indulge in comparative considerations. The line of reasoning is as follows. Firstly, the people far below look the way ants look when viewed more closely. But ants viewed more closely present a visual image which occupies only a small portion of his visual field [we are not told how he knows this, incidentally; presumably some trigonometry and optical theory is involved]. Hence, the people far below must also be presenting a visual image which occupies a small portion of his visual field.

But how was he able to determine in the first place that the people look like ants? If the concept of something "looking like an ant" is to be derived in the way he suggests, it must be subjected to a mass of collateral considerations. Thus, for something to look like an ant viewed from a few meters away [in the relevant respect] is for it to present a particular [yet unspecified] appearance which disposes, or *amounts to* the disposition of, the observer to report that it looks like an ant viewed from a few meters away, *unless*, among other conditions:

Dennett's Response to the Standard Intrapersonal Argument.

Note firstly that although Dennett considers that he has other independent reasons for rejecting qualia [see chapters I and II for my treatment of these] he approaches the inverted spectrum argument at least feigning an open mind on the subject. In other words, he is attempting to evaluate the argument on its own terms as a

---

1. The ant is viewed from a few meters away through a powerful telescope.

2. The ant is submerged in nitric acid and left for a day.

3. The observer reports that something looks like an ant, but is either dishonest, confused or hypnotised, or has been instructed to give a misleading report.

In this example, the epistemological objection would be that it is surely possible for an observer to determine that something looks a certain way [e,g., occupies a small portion of his visual field] without even having to consider any of these collateral conditions. Hence, the [physico-dispositionally topic-neutral] *concept* of something occupying a small portion of the visual field must be logically distinct from the concept of the dispositional traits to which it is supposed to amount. Admittedly, it might turn out through a posteriori investigation that something occupying a small portion of his visual field is *in fact* just his having the said dispositional complex. Nevertheless, it just *is* possible for him to establish that the people far below seem to present a small visual image irrespective of whether or not he has any knowledge of those dispositions. In the same way, we have presupposed, plausibly, that Smith is able to determine in introspection that a property of the type [headache] is occurring without having to consider any extraneous factors of the sort cited by Hill. The crucial question for the inverted spectrum argument is what these properties could amount to in physico-dispositional terms. [See also our discussion of the holistic dispositional analysis of qualia later in the chapter].

purportedly diagnostic test for the dispositional analysis of the concept of a particular quale [or the concept of the experiencing of a particular quale]. In that spirit, he is not entitled to beg the question as to whether the two can be distinguished by introducing his independent reasons for rejecting qualia, and we have argued in any case that his reasons fail to stand up to scrutiny. With this assumed neutrality in mind, then, we can now look at Dennett's interpretation of the standard argument.

From Dennett's point of view the above version of the argument will seem to present a fine example of at least *some* dispositions remaining attached to the experiencing of a particular quale; Smith announces that "this object [qualitatively] looks red to me" just if the object under observation produces Rp [ripe tomatoes did so on Saturday, but grass does so on Sunday]. For a time, at least, he will also report, mistakenly, that red things are green, and vice versa. In other words, Smith's disposition to announce that the object looks [qualitatively] red, and also, at least initially, that it both looks like standard red objects and is red, depends on whether or not the specific quale Rp is produced by it. This appears to support Dennett's claim that Rp is not being imagined to co-obtain with the *complete* state GP, and hence that neither [2] nor [3] has been established by the argument. Clearly, by reformulating the argument with Gp replacing Rp, and GP replacing RP, Dennett will also infer that Gp is not being imagined to co-obtain with the complete state RP, and hence that neither [1] nor [4] has been established either.

The trouble is, however, that although, as Dennett points out, the dispositional state GP in which Smith sees a red object in 2 co-obtains at that stage with his experience of Gp, his state GP is not maintained over time. That is, even if we accept that at stage 2 the quale Smith experiences when looking at paradigmatically red things *is* now Gp and his dispositional state *is* now GP, after a period of adjustment to his knowledge of his newly inverted qualia Smith will report once again, in 4, that ripe tomatoes are red and look like standard red objects. He will undergo at least a partial reversion from GP towards RP while continuing to experience Gp. We can say at least that the state GP is replaced by some other state GP'. Thus, it appears that it is possible to imagine the experiencing of Gp not co-obtaining invariantly with the *complete* dispositional state GP.

In this case, then, we seem entitled to claim at least a modified version of [4]:

[4'] Gp can be imagined to co-obtain with either GP or GP'.

and thence, that qualia are indeed conceptually distinct from any dispositional states whatever.

So it seems that Dennett's insistence that in the envisaged experiment we are obliged to imagine a *complete* reversion of Smith's dispositional state from GP to RP while continuing to experience Gp is unwarranted. All the argument requires is that a *partial* inversion can be imagined to occur, from GP to GP'.

Dennett's treatment of the inverted spectrum argument confirms that he does indeed regard a partial inversion as being inadequate to the cause. Thus, in his own version:

> You wake up one morning to find that the grass has [qualitatively] turned red, the sky yellow, and so forth. No one else notices any color anomalies in the world, so the problem must be in you. You are entitled, it seems, to conclude that you have undergone visual color qualia inversion. How did it happen? It turns out that while you slept, evil neurosurgeons switched all the wires - the neurons - leading from the color-sensitive cone cells in your retinas. [Dennett, 1991, p 391]
>
> What the qualophile needs is a thought experiment that demonstrates that the-way-things-look can be independent of *all* these reactive dispositions. [pp 391-2]

Now there is no question, even in Dennett's mind, that the experiential results of the above unsolicited operation would be noticeable to the subject Smith. The problem picked out by Dennett, however, is that in inverting Smith's spectrum it is difficult to imagine at least some of the reactive dispositions also reversed in the process being restored to their original form, even over time. Thus, whereas he used to find the colour of ripe tomatoes warm, he now finds it cold [because the quale now produced by the observation of red objects, Gp, is of the type which used to be produced by the observation of green objects]. Similarly, the excitable mood formerly induced in Smith under red light is now elicited by

irradiating him with green light. And so on. In such cases, therefore, we might expect Smith's reactive dispositions also to seem inappropriate to the less doctored members of his society. Consequently, at least for Dennett, the experiment is unsuccessful in exposing a conceptual distinction between colour qualia and reactive dispositions. It seems to him that there will be at least *some* dispositions which remain inseparably linked to particular qualia, or to the experiencing of those qualia, throughout the course of the experiment.

Reply to Dennett's Objection.

How are we to respond to Dennett's objection? There are three important points to make.

Firstly, we must point out that even a partial inversion of Smith's reactive dispositions while experiencing Gp is sufficient for our purposes. If Rp is supposed to be conceptually identical with RP, and Gp with GP, it seems clear that a partial inversion will be sufficient. For it is logically impossible that a quale Gp should be imagined to vary independently of *any* of the constituent members of the type [GP], unless the type [Gp] *per se* is at least conceptually distinct from the corresponding dispositional type [GP].

Secondly, it is important to make sure that we are imagining what we think we are imagining. So far we have assumed on intuitive grounds that the scenario envisaged in the inverted spectrum argument really does involve imagining Rp or Gp remaining invariant while the dispositional traits exhibited by Smith change. That intuition, however, could turn out on further analysis to be the result of a conceptual confusion, somewhat akin to the following example.

Consider the case of a sealable gas container. Suppose, for the sake of argument, that physics has established that in fact the pressure P applied on the inner surface of the container is just the rate of molecular momentum exchange M of the enclosed gas in every case. Given that fact, it still might seem possible to imagine a case in which the container walls are under pressure P even though there is no gas inside, and therefore no momentum exchange. If the two events can be conceptually separated in this way, then, the natural

conclusion would be that they are not conceptually identical after all. But what, exactly, has been imagined? Firstly, we think we are imagining a container under pressure. We imagine the walls of the container being subjected to outward forces; they might even begin to bulge a little under the strain. Now, can we imagine that a container under those conditions contains no gas and therefore experiences no molecular momentum exchange? Of course! The container is completely empty; a vacuum. So there we have it. A container whose walls are bulging from the pressure P even though no molecular momentum exchange M is occurring. Hence, P and M must be conceptually distinct.

The problem with this thought experiment, however, is that we are not imagining what we *think* we are imagining. Pressure P is not a bulging effect on the container walls. It is a force exerted at the surface. The bulging walls are merely an *effect* of the pressure. Once this much has been accepted through conceptual analysis, it becomes easier to see at least that pressure P is conceptually identical with the force exerted on the surface after all. The same line of reasoning also applies to the momentum exchange M. How M is produced is irrelevant. Gas molecules rebounding from the surface will do it, but so will other agents. There is multiple instantiation of the *cause* of the momentum exchange M, rather than of M itself. What the molecules do in each case, however, is exchange their momentum on contact with the surface. That exchange of momentum M, or more accurately, the rate of exchange of momentum per unit surface area, just is the force we know as pressure [this is not entirely accurate, but illustrates the relevant point quite well]. Thus, we come to realise through purely conceptual analysis that pressure P and momentum exchange M really are conceptually identical after all. It is only because we began by imagining the *cause* of the pressure on the container [bouncing molecules] and the *effect* of the pressure on the container [bulging walls] that the conceptual identity of P and M was not immediately appreciated. Having thought more carefully about the relevant concepts we now find that it is impossible to imagine a container, gas-filled or not, which undergoes P but not M, or vice versa.

Similarly, then, it might be the case that our intuition to the effect that at least some of the dispositional traits in GP can be imagined not to co-obtain invariantly with the occurrence of Smith's

particular quale Gp is merely the result of another conceptual confusion. For even if the dispositional state GP obtains at stage 2 in the scenario, but dispositional state GP' obtains at stage 4, it might be that it is impossible to imagine what we think we are imagining. Specifically, it might be that when we imagine Smith's dispositions changing from GP to GP' we are *thereby* imagining his quale following suit, because the qualitative change just *is* the dispositional change. To suppose that this is not so might be to *presuppose* the quale we are imagining Smith to experience in 4 is in fact Gp. In any case, the issue which ultimately has to be addressed in this discussion concerns factual identity, not conceptual identity. As we are clearly running into trouble with the conceptual analysis, then, it might now be appropriate to see how the argument explored so far can be adapted to apply more directly to the factual issue.

Thirdly, it might be argued by Dennett that RP and GP are not the appropriate dispositional states after all; that there is some other pair of states which *cannot* be imagined to vary, or vary in fact, in the required way. This point will be taken up later.

A Factual version of the Argument.

The points we have just been considering in response to Dennett's position will now be considered in the context of factual, rather than imaginable, spectral inversion. The first two will be characterised as the claims that, if GP and GP' are assumed to be the only plausible candidates for identity with Gp:

1. A partial inversion, from GP to GP', while Gp obtains, is sufficient to show that quale Gp is *in fact* distinct from any single dispositional state [and this can be extended to other qualia by analogy].

2. An inversion of the sort indicated in 1. can in fact occur.

Let us remind ourselves of the thesis being defended by the qualophile, or qualia-dualist, QD in the present context. It is that qualia occur even though they cannot be fully characterised in terms

of Smith's reactive dispositions. And in order to refute that thesis, Dennett claims:

> What the qualophile needs is a thought experiment that demonstrates that the-way-things-look can be independent of *all* these reactive dispositions. [pp 391-2]

We are now in a position to respond in the following way. Even if the dispositional states GP and GP' really are the only states which might plausibly be supposed to co-obtain [on different occasions] with Gp, Dennett's only possible response is that GP and GP' each amount to instances of Gp in virtue of each having some common property or characteristic in virtue of which each can be said to exhibit Gp. If such a reversion can obtain, then, he must concede that neither GP nor GP' *per se* constitutes the unique dispositional characterisation of Gp. He must claim instead that GP and GP' share some narrower set of dispositional traits Gx in virtue of which each is to count as an occurrence of Gp. We can then refer to the proposed set Gx as the *diagnostic* set for Gp; no dispositional set can exhibit or amount to an instance of Gp unless it has the complete set Gx. Let us suppose, then, that there is some set Gx such that each of the constituent members of Gx is *necessary* for the occurrence of Gp. To suppose that there is no such set *whatever* would amount to conceding that the diagnostic set Gx can itself be multiply instantiated by completely distinct sets of dispositional components.[27] In parallel fashion, we can suppose that there is also some narrower set of dispositional traits Rx which RP and RP' must have in order to count as occurrences of Rp. If the identity thesis is to have any meaning, then, we must assume that, construed as universals, Gx is held to be identical with Gp, and Rx is held to be identical with Rp.

The point we would want to make in reply to this latest suggestion is exactly parallel to the objection already raised. The "qualophile", surely, still has much less to do than Dennett appears to suppose. Dennett must claim that an occurrent quale is of the

---

27 We argue later in the chapter that multiple instantiation of this sort precludes the possibility of an identity relation.

type [Rp] just in case it contains all the members of Rx. Similarly, he must claim that an occurrent quale is of the type [Gp] just in case it contains all the members of Gx. Now, since Rx has been specified as the dispositional state comprising all the necessary constituents of Rp, it follows that there must not be even *one* constituent of Rx which need not obtain in the event of an occurrence of Rp [and similarly for Gp and Gx]. So the change from Rp occurring to Gp occurring must be accompanied by [or amount to] a change from having the complete set Rx to having the complete set Gx. Hence, we can infer that in order to succeed the inverted spectrum argument need only show that there is no set Rx, or Gx, of the sort required. But we can show that this is indeed the case if just *one* disposition D which would have to be a constituent of Rx, or Gx, can be shown not to co-obtain invariably with Rp, or Gp.

Admittedly, since Rx and Gx have yet to be specified, we must concede to Dennett that at least all of the *plausible* candidates for membership of Rx and Gx respectively must be shown not to co-obtain invariably with Rp and Gp. Even so, we can see that this indeed appears to be the case; for even the most likely candidates for membership of Rx might also occur in the presence of Gx. Thus, for example, Smith's occurrent quale might change from Rp to Gp even while he retains the disposition to report that he is seeing red [this would depend on how much collateral information he has, and on whether he was confused, dishonest or hypnotised, or in some other way adversely motivated], or even that he is *experiencing Rp*. It seems clear that there is no conceivable dispositional trait which must invariably co-obtain with Rp after all. Since this much seems clear, then, the inverted spectrum argument appears to succeed. What we are entitled to conclude is that Dennett's demand that *all* of the dispositional traits belonging to Rx will survive the change in experience is simply false. If Rx really is the set of dispositions which constitutes Rp, [and we can allow that Rx comprises any individual dispositions he cares to think of], we need only find a single feature of that set to be wrongly included in order to infer that the supposedly diagnostic set Rx is incorrect. The most he can claim in this respect is that since he has not *specified* which traits are members of Rx, we will have to consider all the plausible candidates in order to reach our conclusion.

There is no possibility when the offending dispositional traits are relatively simple that our conclusion has been drawn on the basis of a conceptual confusion. For it is quite clear that even if we concede that what we have been referring to as Rp is just Smith's *seeming* to experience Rp, the envisaged possibilities just do occur. Dennett suggests to the contrary that we can all begin to see his point of view with the help of a basic piece of philosophical equipment; an ordinary pair of army surplus infrared sniper's goggles. Using these, he claims, it is possible to witness first-hand the phenomenon of pre-experiential adaptation. Thus, whereas everything appears at first in "weird and hard-to-distinguish colours" [p 394], there comes a time when you have adjusted sufficiently to the effect of the goggles to render the appearance of colours relatively normal again. Couldn't it also be the case, he wonders, that someone with colour-inverting lenses might similarly adapt until his qualia had *completely* reverted to their original state? If it is the case, and it must be the case for Dennett because the dispositional adjustment is identical with the qualitative adjustment, we will be forced to concede that Smith's transition from Gx to Rx must be accompanied by a corresponding change in the occurrent quale from Gp to Rp, and therefore that the two cannot vary independently of one another. Perhaps Dennett is right, and the inverted spectrum argument is based on unsubstantiated speculation or confused thinking. In terms of our example, he would have to maintain that the deviation from Gx cannot obtain while Gp remains constant. If it did, there would be two distinct dispositional sets to be identified with a single quale. Hence, he must maintain that *any* deviation from Gx must be accompanied by a qualitative change, and that in consequence there can be no deviation of the sort needed to show that his dispositional thesis is false.

Is this position sustainable? Even Dennett's infrared goggles cannot help him here. For even if it turns out to be true that, as a result of "pre-experiential adjustment" over time, Smith's qualia will eventually revert to the original state Rx when he is seeing a red object through the goggles, he quite clearly needs time and, perhaps, collateral information for this to occur. There can be no question that on putting on the goggles he will initially have, or *seem* to have, a Weird-p experience. But there can also be no question that, on learning that the experiential change has been

induced by the goggles, he will also be able to revert quite readily even to some of the dispositions which Rx, uniquely, should be expected to contain, without his weird quale being affected in any way. On realising how the goggles are affecting his vision, for example, he might quickly learn to *report red objects as being red* even though he continues to experience them weirdly. We can even envisage situations in which he would in fact be disposed to report that "*I seem to be experiencing Rp*" even though he does not seem to be doing so [he might be permanently confused, dishonest or hypnotised, for example] and, hence, situations in which no amount of pre-experiential adjustment over time will remove that disposition. Thus, although his report suggests that he has the Rp experience, there are possible circumstances in which he will, in fact, continue even indefinitely to make the misleading report when he in fact has some other experience. The crucial point is that even the dispositional traits which Dennett must presumably regard as being essential constituents of the appropriate set Rx can be exhibited while Weird-p is being experienced.

Furthermore, the report of a qualitative change in experience is no less real by dint of Smith's unreliable memory. For if it makes sense to say in the first place that Rp obtains, where "Rp" refers to a particular [intelligible, but intentionally inexistent] experiential quality, and similarly in the case of Gp, then it also makes sense to say that the occurrent quale changes from one to the other, irrespective of whether or not he remembers reliably the former qualitative character of Rp [just as, in Chapter II, we saw that Smith's inability in the colour-phi experiment to decide between two versions of how his qualitative experiences might have proceeded fails to indicate that he had no such experiences]. That is enough to show that one or more of the reactive dispositions which Dennett regards as essential components of the set comprising a particular quale is not an essential component after all, even as an a posteriori matter of fact. The suggestion that Smith's memory is unreliable simply *adds* to the possible variations in the dispositions he might exhibit while experiencing a particular quale.[28]

28. Ned Block thinks that Dennett's memory-based objection is circumvented by reconstruing the inverted spectrum argument in terms of an *inverted earth* version [Block, 1990]. Instead of arranging for Smith's spectrum to be inverted, we arrange for it to remain constant in the face of inverted reactive dispositions. This way, Block believes, there can be no doubt that the two are separable.

Smith has a lens transplant, as before, but is transported under anaesthetic to an inverted world, where everything is the wrong colour. The sky is yellow, the grass red, and so on. To add to the confusion, however, the inhabitants use inverted colour terms. So for them as well as for Smith the sky is described as being "blue" and the grass as being "green". In terms of intentional content, then, we can say that for both Smith and the locals the intentional content [or intentionally inexistent referential domain] of "red", for example, is the same. In terms of qualia-beliefs, the envisaged sequence of events over the weekend might be outlined as follows.

> Stage 1. [Saturday Morning] Smith exhibits normal dispositions and qualia-beliefs.
>
> Stage 2. [Saturday Night] Quale-inverting lenses installed. Smith transported to Inverted Earth.
>
> Stage 3. [Sunday Morning] Qualia-beliefs are inverted, so Smith *notices no changes* [yellow sky produces the Bp-belief].

Dispositionally, Block claims, he would remain unchanged; Smith reports that the sky is "blue, as usual". The difference is that *he is now wrong*. The sky is in fact yellow but he has the quale-belief produced on Saturday by a *blue* sky. Now, one reason why this version might seem irrelevant to the present debate is that it characterises Smith's dispositional state only in terms of the *representational* content of Smith's beliefs and dispositions; his attendant *physical state* might still be BP at stage 3. Hence, the two might still be inseparable. We shall overlook this point in the present context, however, since we are interested here in the problem of memory malfunction. The advantage of Block's version in this respect is supposed to be that there is *no internal disturbance for Smith*

28 [continued].

during the transition. Everything seems to be the same as before, even to the extent that the same objects seem to retain the same colour throughout. The changes are all *external*, so there can be no problem of indeterminacy or fallibility concerning Smith's memory.

The problem with Block's account is that the so-called advantage is an illusion. The original advantage of the *intrapersonal* inverted spectrum hypothesis was the availability of Smith's introspective report at stage 3, when he reports that the qualitative content of seeing red is the same as that of seeing green used to be. Dennett accuses him of memory malfunction or even, with Rey [1991], fails to understand what he is saying. But how does the Inverted Earth hypothesis remove the memory-malfunction objection? When Smith has his lens implant and arrives on Inverted Earth Block claims that Smith notices no difference in the qualitative content of his experiences or qualia-beliefs. But the assumptions Block requires to support this intuition are just those required by the inverted spectrum advocate to support his claim that the experimental subject will experience *inverted* qualitative content. If this is correct, therefore, Block's example is no more or less convincing than the original intrapersonal inverted spectrum argument. Block says;

> In the latter case [intrapersonal inverted spectrum] the subject's internal disturbance renders his first person reports vulnerable to doubt. But you, the subject of the Inverted Earth case, have had no internal disturbance. Your move to Inverted Earth was accomplished without you noticing it - there was no period of confusion or adaptation". [1990, p 65.]

But it is not at all clear that his example does eliminate the uncertainty with regard to memory. Suppose that in each of the two worlds there are two types of object; trees, which are green on Earth and red on Inverted Earth, and ripe tomatoes, which are red on Earth and green on Inverted Earth. The advantage with Block's experiment is supposed to be that because Smith has inverting lenses fitted before waking up on Inverted Earth, he continues to experience trees and tomatoes just as before. But now consider the *standard* intrapersonal inverted spectrum hypothesis. Here, Smith wakes up to find that trees now look the way ripe tomatoes used to

Thus, if Gx is *any* set proposed as the diagnostic dispositional set for Gp, the original statement that:

1. A partial inversion, from GP to GP', while Gp obtains, is sufficient to show that quale Gp is *in fact* distinct from any single dispositional state [and this can be extended to other qualia by analogy].

now becomes:

---

look, and vice versa. But since the physiology of seeing red things without the lenses is the same as the physiology of seeing green things with the lenses, it follows that Smith must experience no qualitative change from the experience of the colour of trees before the lens implant to the experience of the colour of ripe tomatoes after the implant. The certainty which Block hopes to introduce by formulating the Inverted Earth hypothesis is thus founded, not on memory-based considerations at all, but rather on the principle that identical physiology entails identical experience. But if this principle were sound, it would also apply equally effectively to the original inverted spectrum hypothesis. Thus, the fact that seeing red objects is now like seeing green objects was yesterday could be determined from the fact that red objects now produce the physiological state which green objects used to produce. The opportunity for memory malfunction, however, is the same in both the inverted spectrum case and the Inverted Earth case and therefore either constitutes a valid objection to both or neither. In the Inverted Earth case Smith has to remember that seeing ripe tomatoes yesterday was like seeing ripe tomatoes today, whereas in the inverted spectrum case he has to remember that seeing ripe tomatoes yesterday was like seeing trees today, and vice versa. The only difference is that the experience of each qualitative type now comes from seeing a different type of object. So if Dennett's memory-based objection were problematic, which it is not, Block's proposal would fail to circumvent it.

1'. Any deviation from Gx, while Gp obtains, is sufficient to show that quale Gp is in fact distinct from any dispositional states [and this can be extended to other qualia by analogy].

And we have seen that for any set Gx such deviations can occur in fact, even if the property Gp is taken to be merely the intelligible object of an occurrent *belief*. In the latter case, we can say that we can undergo a deviation from any set Gx while retaining the intelligible belief that we are experiencing Gp. Hence, having that intelligible belief cannot amount to being in any particular dispositional state either.

Thus, for Dennett, there must be at least some dispositional trait which cannot change at all without a corresponding change in the associated quale, or the corresponding experiential belief. If that were so it would follow that we are unable to distinguish the quale [or belief] from that dispositional trait in virtue of any spectral inversion. Once we accept that this has to be his line of argument, however, it appears that there are no suitable candidates available.

Finally, the question arises as to whether the dispositional account of qualia might be saved by construing what we take to be the experiencing of particular qualia as episodes of seeming to do so, in some other sense. Thus, if qualia are supposed to be the occurrent qualities of experience, it might be suggested that there only *seem* to be such qualities, or even that no experience occurs. Consider, then, the blanket proposal that we only seem to have any experience at all, including the experience of qualia. This strategy will not help, however. For even if we construe the having of a headache thus, as merely seeming to have an experience of the type [headache], it remains true that we are able to determine specifically that we at least seem to experience a headache. Thus, whenever we have been referring to a headache, or to quale Rp, we should have referred instead to the event of seeming to experience an item of that type. No discriminatory power is lost, but the object of that discrimination has been recast. But that just means that the dispositional theorist is now obliged to defend his position against the inverted spectrum argument in an exactly parallel fashion, with exactly parallel results. Whatever occurs when we seem to experience a headache, or seem to experience quale

Rp, no purely dispositional account of the sort we have been considering can do it justice.

First Response; The Holistic Approach to Dispositional Analysis

Originally, we attempted to provide an account of reactive dispositions solely in terms of Smith's behavioural responses to unspecified input. Thus, we construed the reactive dispositions which are to be identified with Rp in terms of Smith's disposition to make such reports as "that is red", or "that looks red to me", irrespective of the actual colour of the object prompting the response or the collateral information available to him. Similarly, his disposition to eat a tomato when it produced Rp was specified irrespective of his knowledge of the actual colour of the object being viewed. We now see, however, that these additional considerations must be built into the account after all. An occurrence which elicits the response "that is red" only counts as an occurrence of Rp if, among other conditions, it is not held by a Smith who knows that he is viewing a red object through colour inverting lenses [in at least some such cases he would experience, or seem to experience, an occurrence of Gpl. Again, to use one of Dennett's own examples, [p 391], he will only be disposed to pass the ball to the players in red [his own team wears green] so long as he does not know that he is wearing colour inverting lenses. His disposition to do so will belong to *GP* only if the object involved is red and he does not know about the lenses he is wearing.

Our major difficulty with this result is that the collateral information becomes an integral part of the dispositional account of qualia. One of the components of GP is the disposition to;

1. pass the ball to the players in red while

2. wearing colour-inverting lenses but

3. being unaware of their effect.

Similarly, a state of affairs which elicits Smith's report of an occurrent instance of Rp will only count as a genuine instance of Rp if he is neither dishonest, confused nor hypnotised and has not been

instructed to report the opposite of his actual experience [or seeming experience] while viewing a green object in standard conditions.

Clearly, the possible range of conditions and states of knowledge in which Smith's dispositions would have to be specified is extensive, if not infinite. We would be obliged to determine all possible combinations of viewing conditions and states of knowledge in which he could indicate by his responses, which would also have to be specified, that an instance of Rp is occurring. Reverting to an earlier example, since Smith is evidently able to determine that he has a headache, he is able to determine that an occurrent item is of the type [headache]. Since, epistemically, this might be assumed to be a single type [we can be more specific in order to arrive at a single type if the epistemic facts permit], it is impossible that the type [headache] is identical with each of a number of distinct dispositional types, [HP], [HP'], etc. There must be at least some common dispositional set Hx in virtue of which the distinct dispositional types can be said to share the common property of constituting the single type [headache]. Each of the component types then only exhibits an instance of the type [headache] if it contains the complete set Hx. Clearly, we cannot say just that they share the common characteristic of being the type [headache] without begging the question as to what constitutes being a headache. But, taken in isolation, there is no other dispositional respect in which the component types resemble one another in the required way. To accommodate the above considerations the identity statement must therefore be modified accordingly. Thus, if we take as an example the disposition to report an object as being green, an occurrence of GP must now *include* Smith's reactive disposition to:

1. Report a green object as being green, unless spectrum inverting lenses are worn unwittingly.

2. Report a red object as being green if spectrum inverting lenses are worn unwittingly.

3. Say nothing if he is mute, or something else entirely if he does not speak English.

In other words, the various reactive dispositions turn out to be facets of a single dispositional set if this is expanded to encompass the possible dispositional sets in which Gp might obtain. Smith's dispositional set which constitutes the occurrence of Gp is such that if he does not wear inverting lenses he will make a report of a certain type and if he does wear inverting lenses without knowing it he will make a report of a different type. The disposition to react in a certain way when wearing inverting lenses unwittingly was there even before he wore them. This, then, heralds the beginnings of context dependence for the dispositional thesis. An occurrence of Gp is identical with a certain dispositional set only if this set is expanded to incorporate each of the individual dispositional sets in which Smith's Gp would occur.

Such a dispositional set will be disjunctive, and in view of the large number of possible member sets, we might also refer to the set as *holistic*. A particular quale Rp will then be characterised as the holistic set which incorporates all of the individual sets [Rx.....Rx^N], any one of which would amount to an occurrent instance of Rp. Such a characterisation of any quale might seem preposterous. Thus, while we might readily concede that an occurrent Rp would *generate* the holistic set, it is quite another matter to accept that Rp is identical with that set. It is clear for present purposes, however, that the inverted spectrum argument in any form is powerless to repudiate such a proposal. For if *all* of the individual dispositional sets Rx.....Rx^N are included in the holistic and disjunctive set, there can be no possibility that the latter might not obtain even though Rp does [even if Smith's dispositions change from one occurrence of Rp to another *over time*, while all other factors remain invariant, the time of the individual occurrences can be built into the dispositional account].

Furthermore, we can also see that the epistemic considerations explored in chapter III are of no avail. In terms of a dispositional analysis of qualia, the qualia-dualist QD is claiming that:

> [QD] A quale of the type [headache], or [Quale Hp], is an occurrent irreducibly non-dispositional item.

Thus, while QD agrees with QP that the epistemic situation Eq occurs, he claims that the quale Hp identified in that situation is

an occurrent but irreducibly non-dispositional item. We can begin to see what this claim amounts to by dividing QD's position into two distinct components. Thus, he claims that:

> [QD]1 The type [Quale Hp] is *epistemically* topic-neutral with respect to any dispositional type [HP], and in addition, even to the general type [Dispositional set].

In this claim he is in agreement with QP. Thus, in regard to [QD]1, whatever the fundamental constitution of a headache Hp might be, both theorists are committed to the evident fact that it is possible in Ep to determine that they have a headache without being able to determine either that it is a dispositional set of any type [HP] or even that it is an item of the type [Dispositional set]. But this would be possible even if physicalism were true, just because the occurrent epistemic state Eq is such that in Eq a complete understanding of the physical account of experience, including an understanding of physical *types*, is not available. Where QD disagrees essentially with QP, however, is in claiming that:

> [QD]2 Qualia are occurrent [intentionally inexistent] items of a non-dispositional type (e,g., [Quale Hp]).

Clearly, he cannot support this claim by referring to [QD]1, since [QD]1 is compatible with Hp *in fact* being a dispositional set HP, as QD concedes [and our brand of physicalism dictates that even *types* are to be identified]. The question is whether the inverted spectrum argument can provide the further support he needs.

Firstly, the inverted spectrum argument can at least be employed to establish that there is a sense in which we can *imagine* the holistic dispositional set varying while a quale Hp remains invariant. For since the type [Quale Hp] is evidently topic-neutral with respect to any particular dispositional type [HP], we might legitimately infer that it is possible to imagine an item x being of the first type but not the second. Thus, we might identify x as being of the type [quale Hp] without knowing anything at all about the type of dispositional state a subject might be in when x occurs. In that case it would be possible to imagine that x, as an item of type [Hp], should turn out not to co-obtain with any particular holistic dispositional set HP. If Hp and HP are in fact identical, we could

then infer that they are at least *conceptually* distinct, since our concept of x is nothing more that the topic-neutral concept of *whatever dispositional set Hp happens to be.*

The trouble is, however, that QD will be unable to use the inverted spectrum possibility to substantiate his position at the expense of [QP]. For if it is possible to imagine the required variation *whether or not Hp is in fact identical with HP*, it follows that this possibility cannot establish the non-dispositional nature of Hp. At the factual level it remains an a posteriori possibility that Hp should *not* vary against a fixed backdrop of the entire holistic dispositional set. Hence, at the factual level, the inverted spectrum argument is unable to establish that an holistic dispositional analysis of qualia is false, without establishing by a posteriori investigation that there is no appropriate dispositional set. Indeed, if our previous argument is sound, it appears that in principle a complex and disjunctive dispositional set can always be contrived to accommodate all of the dispositional variations exhibited by Smith while experiencing Hp.

Second Response; Disjunctions and Multiple Instantiation.

If the holistic dispositional analysis of qualia is thought to be implausible, however, there is another possible strategy available to QP. The dispositional thesis in question entails that Hp can occur only if a particular set of dispositional traits HP is exhibited. If, ex hypothesi, disposition HP_1 is any constituent of the diagnostic set HP for the quale Hp, then, it is impossible for that quale to occur in the absence of HP_1. But it might be suggested that the appropriate diagnostic set is disjunctive, so that *either* HP or some other set HP' must obtain for Hp to occur. Hence, our discovery that there are no particular constituent dispositions which must be exhibited can be accommodated. At the very least, however, this suggestion implies that we have a case of multiple [dispositional] instantiation of a single type of experience. Several alternative sets of dispositional traits can each amount to an instance of the type [Hp].

But this entails that *neither* [HP] nor [HP'] is identical with [Hp]. To take an example from music, it is a fact that a foursome must be

a group of four individuals, but not that it must be a group of four musicians [a quartet]. The type of individuals involved can be altered without affecting the fact that a foursome is present. All quartets are foursomes, but not all foursomes are quartets. A fourman bobsleigh team would also be a foursome. Now, since the constituent members of the type [foursome] can be of various types - people who may or may not be musicians - the type [foursome] cannot be *identified* with the type [group of four musicians]. In other words, it is only by drawing the category of a constituent member of the type [foursome] too narrowly for an identity relation to obtain with that group that we have been able to create the possibility of multiple instantiation of types. The referent of "foursome" can be identical with the referent of *either* "group of four musicians" or "fourman bobsleigh team" in any particular instance, which amounts to the multiple type-instantiation of the type [foursome].

The problem can be expressed more formally in the following way. As before, we shall assume for convenience that *occurrent qualia* are the subject of the discussion. *Experiencing* qualia and merely *seeming* to do so will then present parallel problems. As before, the PPD candidates for identity with qualia will taken to be the overtly dispositional traits exhibited by the subject; i.e, observable behavioural responses to sensory input and collateral information about the external state of affairs.

Suppose firstly that there are two distinct dispositional state types [HP] and [HP'] each of which is an instance of [Hp]. We can then argue that since:

1. [HP] is identical with [Hp].

and:

2. [HP] is distinct from [HP'].

it follows logically that:

3. [HP'] is distinct from [Hp].

And from this it follows by *reductio ad absurdum*, after re-running the above argument with [HP] and [HP'] interchanged, that [Hp]

cannot be identical with either [HP] or [HP']. And since this must be true when [HP] and [HP'] are any two dispositional state types whatever, it follows that multiple instantiation [i.e., the identity of [Hp] with any two distinct state types] is logically impossible. Hence, if ex hypothesis the dispositional facts are the only facts available, type [Hp] in 3 must be identical with some other dispositional type which always accompanies, or is contained in, both [HP] and [HP']. But we have seen that there is no such type.

We have seen that no relatively simple dispositional type [HP] can be identical with [Hp], since none has been found to be invariantly associated with [Hp]. So [Hp] is not identical with any relatively simple dispositional type. The inference, then, must be that [HP] can only be a more complex dispositional type. But if no relatively simple dispositional type is invariably associated with [Hp], then *a fortiori* no more complex single type can be invariably associated with [Hp] either. So we may infer that no single dispositional type *whatever* is invariably associated with [Hp], or any other quale type. And this suggests that the multiple instantiation of qualia types can only be accommodated by resorting to the holistic approach already discussed. While, on the assumption that the qualia as supposedly discerned in introspection are intelligible, this is an intelligible suggestion it is difficult to see how to corroborate the suggestion, as we have already argued.

The Distinction between *Having* Dispositions and *Exhibiting* them.

Reverting to the case of colour perception, the physicalist might be able to salvage his position by suggesting that what RP and RP' have in common, in virtue of which each constitutes an occurrence of Rp for Smith, is just that *in standard conditions* [which would have to be specified] Smith *would* have reported an occurrence of Rp and exhibited any other dispositions which might be thought appropriate to having an occurrence of Rp. We can then say that the common characteristic Rx of RP and RP' is such that:

Rx[defn.] Smith's dispositional state has Rx just if, in standard conditions, Smith *would* have exhibited Rx-appropriate dispositions.

We could then go on to explain that the conditions in RP are standard, but that those in RP' are non-standard. Suppose, then that in RP', but not in RP, Smith is just dishonest. From our foregoing analysis, this proposal would appear to be at least consistent with the facts, and it also explains how Rp can occur in RP *and* RP'. Thus, if:

1. RP constitutes the set of Rp-appropriate dispositions,
2. RP' constitutes a set of dispositions which *would have been replaced by RP* had Smith been honest.

it follows that RP' differs from the occurrence of Rp in standard conditions only in respect of Smith being dishonest, and therefore in failing to exhibit the Rp-appropriate dispositions. Clearly, we must accept that there will be epistemic situations in which 2 is true even according to the dispositional thesis. Smith might have a headache, for example, and be able to determine that he does so, even though he does not want to talk about it or reveal it in any other way. Hence, the dispositional account of what constitutes Smith's headache is to be distinguished from the behaviour he *actually exhibits*.

It seems that this distinction between the dispositions exhibited by Smith and those he *would* exhibit in standard conditions offers the physicalist a much stronger position. Thus, it surely makes sense to suggest that Smith has a headache just if he would exhibit certain characteristic dispositions in the absence of extenuating circumstances. The physicalist can then explain what it amounts to for Smith to have an occurrent Rp in any conditions in the following way.

> Rp[defn]. Smith has an occurrent Rp just if he is in a neural state N which, in standard conditions, would lead him to exhibit Rp-appropriate dispositions.

Here, the type [Rp] is characterised dispositionally, and the reference to neural state N is a topic-neutral reference to whatever neural state or states would produce the appropriate dispositions in standard conditions. Thus, the need to resort to a disjunctive or holistic dispositional characterisation of Rp can be avoided.

Further, the possibility of there being several distinct neural states each of which satisfies the condition for N is now acceptable, since an occurrence of Rp amounts just to the occurrence of *any* neural state which would produce the appropriate dispositions in standard conditions. Nevertheless, the above definition does imply that there is some neural state N [or disjunctive set of states] which, in standard conditions, would invariably produce Rp-appropriate dispositions. If this were not the case, it would not make sense to say that N *would have produced* those dispositions.

The inverted spectrum argument can now be employed to challenge the above account in the following way. Since the reference to neural states is topic-neutral in the above account, we cannot claim that the specific neural states referred to fail to satisfy the specified dispositional conditions. What we can claim, however, is that there *are* no specific neural states which satisfy the specified dispositional conditions. For any proposed neural state N, it is possible that Smith in N would *not* exhibit Rp-appropriate dispositions in standard conditions. Thus, while the inverted spectrum advocate might concede that:

1. An occurrence of Rp would invariably be accompanied by the exhibition of Rp-appropriate dispositions in standard conditions.

he need only establish that:

2. There is no neural state N [or disjunctive set of neural states, N_1 ... N_N] which would invariably be accompanied by the exhibition of Rp-appropriate dispositions in standard conditions.

in order to infer that Rp is not neurally constituted.

Once again, however, it seems obvious that since Smith evidently *does* exhibit dispositions of one sort or another in standard conditions, it must be trivially true that on any particular occasion there is at least some neural state or other which disposes him to do so. If that is the case, it will be impossible for the inverted spectrum argument to demonstrate otherwise, since there need be no further specification of the neural state involved. The

outstanding problem is that we still have no idea whether the identity thesis is true; whether an occurrent quale just is a neural state of the type thus characterised in dispositional terms. Within the context of our current, limited neurophysiological knowledge, it remains only an a posteriori possibility that qualia should not be neurally constituted in the proposed way.

Conclusion

The aim of the original inverted spectrum argument is to establish that qualia are at least conceptually distinct from any behavioural or reactive dispositions whatever. In order to establish further that qualia are *in fact* distinct from such dispositions, then, some additional argument would have to be produced. We find that, in the absence of any further argument, it is impossible to infer the factual case from the conceptual [Kripke offers such an argument, as we shall see in a subsequent chapter]. Even if Smith's headache or quale Hp is dispositionally constituted, he can determine that he has a headache, or a quale Hp, without even knowing that it is dispositionally constituted. Hence, his ability to categorise an item as a headache or quale Hp without also being able to categorise it as any dispositional type whatever implies that there is a conceptual distinction between the respective types. It is *this* conceptual distinction which enables him to imagine an item being of the type [Quale Hp] but not being of any dispositional type whatever. Since he can do this even if headaches and qualia in general *are* dispositional states, then, we may not infer from his imaginative skills that the two are distinct in fact. In order to draw that inference, we would need to establish that the quale in question can *in fact* obtain in the absence of any plausible dispositional candidate.

Taking dispositions of any relatively simple type as candidates for identity, it appears to be a fact that Smith can have a quale of a particular type even though no disposition of that type invariably co-obtains with it. It follows from this that no disposition of a more complex but non-disjunctive type co-obtains invariably with it either. Hence, the only remaining possibility is that the dispositional candidates are either of holistic and disjunctive

types or the quale can be multiply instantiated by any one of a number of dispositions of distinct types.

That they should be identical with holistic and disjunctive dispositions seems implausible, but the inverted spectrum hypothesis offers no obvious way of determining that the identity does not obtain. For if the disjunctive disposition incorporates all of the individual dispositional traits which might co-obtain on any particular occasion with Smith's headache, it seems to follow that it is impossible to show that there are occasions on which he has a headache but none of those traits obtains. Even if it does follow, however, this does not imply that the dispositional set is *identical* with the quale in question, since it remains plausible to suppose just that the experience of a quale Q merely *generates* the specified dispositional complex. Hence, the most we can say is that the credentials of such an identity thesis remain completely *immune* from an argument of this form.

The alternative proposal, however, that a quale of a particular type should be multiply instantiated by dispositional or neural states of distinct types on different occasions, is a proposition which we find logically incoherent. For if [Rp] is identical with [RP], or alternatively with [N], it is not possible that it should also be identical with some type [RP'] or [N']. Nevertheless, it remains an a posteriori possibility that Rp should turn out to be identical with a *single* neural type [N], no matter how complex.

Finally, we suggested a way of circumventing the problems encountered in the dispositional analysis. [Rp] might be just a type of neural state [N], such that instances of [N] would invariably be accompanied by Rp-appropriate dispositions *in standard conditions*. The advantage of this version is that multiple instantiation is no longer a problem. For we can say that an occurrent quale Q might be *any* neural state which *would* generate Q-appropriate dispositions in standard conditions, and there might be several different types of neural state which would satisfy this condition. The important point is that there must surely be some such neural state or states; for if the neural state *and* the standard conditions are specified, there appears to be nothing else to specify. So since there are occasions on which Smith *does* exhibit Q-appropriate dispositions in standard conditions, it must be trivially true that the appropriate neural

states occur. Hence, the inverted spectrum argument is powerless to show that they might not. Whether this analysis of qualia is correct, however, would again have to be established by a posteriori investigation. The crucial question is whether an occurrent qualia is *identical* with an occurrent neural state of the specified type. In order to answer this question, some further investigation will therefore be required.

What is needed, therefore, is some further argument to establish that qualia cannot be characterised in terms of either holistic and disjunctive dispositions, or neural states of any particular type [even if characterised in terms of the dispositions they would generate in standard conditions]. The argument we are about to consider sets out to do just this. The knowledge argument purports to show that someone can know everything of a physico-dispositional nature and yet not know the distinctive character of qualia. If it succeeds, then, the conclusion will be that the character of qualia is non-physico-dispositional in nature, and hence that the qualia-dualist's thesis [QD] has been vindicated.

What we must acknowledge at this stage is that the so-called "qualitative character of an experience", and even "experience" *per se*, has yet to be specified intelligibly as an occurrent phenomenon in any sense which has been shown to be incompatible with physicalism. And the difficulties cited for the inverted spectrum argument have been shown to apply equally to qualia conceived as the mere objects of intelligible beliefs. While [QD] enjoys the advantage of being an *intelligible* thesis [that physicalism is incomplete] even though the omissions have yet to be understood, then, we can only establish that it is true if physicalism can be shown to suffer from genuine omissions. In order to do so, it might turn out that the purported omissions must still be intelligibly specified in such a way as to secure the required conclusion.

Chapter V

REDUCTIVE PHYSICALISM AND THE KNOWLEDGE ARGUMENT.

Introduction

We have established so far that the qualia-dualist [QD] is committed to the following theses.

[QD]1. Qualia are [intentionally inexistent] irreducibly non-PPD properties.

[QD]2. Qualia are occurrent properties.

And hence:

[QD]3. Qualia are occurrent irreducibly non-PPD properties.

But the inverted spectrum possibility fails to preclude the a posteriori reduction of our qualia to neurophysiological traits characterised in dispositional terms. If dispositions are construed in terms of the behaviour Smith *would* exhibit in standard conditions - however those conditions are specified - it becomes trivially true for the physicalist that there must be some occurrent neurophysiological state of Smith in virtue of which he would do so. Thus, if physicalism is true, having an experience Rp [a type-[Rp] experience], for example, will amount to being in some such neurophysiological state. And since Rp is thus characterised in terms of a particular type of *dispositional* set [RP]; that is, in terms of the dispositional property RP, it follows that *any* neurophysiological state which would lead him to exhibit RP in standard conditions will constitute an occurrence of Rp. Type-identity is thus dispositionally construed, while constitution is construed, topic-neutrally, in neurophysiological terms.

The inverted spectrum argument is powerless to repudiate this form of reductive physicalism. For since Smith evidently does exhibit Rp-appropriate dispositions on occasions, and in standard conditions, the physicalist's claim will then be [trivially] that he is in some

neurophysiological state or other on that occasion. His thesis will then be that he exhibits those dispositions *in virtue of* being in some such state N, without having to specify which sort of state that is. It is then irrelevant that in standard conditions Smith might exhibit RP on different occasions while being in quite distinct states N and N', since the physicalist has not specified N as being of any particular type. N is just *any* state in virtue of which RP would be exhibited in standard conditions.

The only relevant rejoinder for the inverted spectrum advocate is then that for any particular neural state [type] N which *does* engender Rp-appropriate dispositions on one occasion in standard conditions, N might not do so on another occasion in the same conditions. For only then would it make no sense to say that having an occurrent Rp amounts to being in some neural state or other which *would*, in standard conditions, engender Rp-appropriate dispositions. There would be some N for which this would not invariantly be the case, and hence Rp, characterised dispositionally, would not be neurally constituted. Since we have stipulated "standard conditions" in this account, however, it is difficult to see straight off that the physicalist's claim thus formulated is false. Further a posteriori investigation would be required to establish that not all occurrences of Rp supervene on a specific disjunctive neural set [N... N_N *in standard conditions. Even if they did, however, the physicalist's thesis would also* remain uncorroborated. For the existence of a rigid supervenience of this sort would still not entail that Rp is neurally constituted.[29] The inverted spectrum argument has nothing further to offer in this matter. The outstanding question for QD, then, is still whether Rp *is* neurophysiologically constituted.

Using a form of knowledge argument, we now attempt to show that no such reduction of qualia is correct and hence that [QD]3 is true. Again, we shall assume at the outset that the physicalist and QD are

---

29. For a concise but clear evaluation of the concept of supervenience, and its shortcomings in the present context, see Paul Snowdon's, "On Formulating Materialism and Dualism".

agreed on the contents of the set [S] of occurrent physico-dispositional items, *as acknowledged by current physical theory*. The reductive thesis will then be that any occurrence of Rp amounts to nothing more than the occurrence of a state which is both constituted by, and characterised in terms of, the members of [S].[30]

The Standard Knowledge Argument.

In the standard argument, the charge being levelled against the physicalist thesis [QP] is that it fails to make sense of or take into account the intrinsic and non-physico-dispositional properties of experience, such as the phenomenal properties we refer to as "qualia". The knowledge argument has been employed widely in the literature in an attempt to refute physico-dispositional accounts in general [functionalism in particular], and it might appear that our discussion fails to draw a sufficiently clear distinction between the various positions. The important point here, however, is that although the physicalist's account might claim an identity relation between qualia types and functional types, for example, the physicalist's assumption must be that items of any functional type are themselves grounded in physically constituted token states. Thus, even though functional types might be topic-neutral with respect to neural or physiological types, for example, it remains true that an item of a functional type must be fundamentally physical in constitution. Hence, any refutation of physicalism per se will constitute a refutation of all physico-dispositional accounts too.

In his version of the knowledge argument, Frank Jackson conducts an imaginary experiment in which Mary, a brilliant neurophysiologist, is born and raised to adulthood in a completely colour-free room.

---

30. In order to circumvent the problem of distinguishing intelligibly between [QE] and [QR] the reductive thesis is being construed just as the thesis that all *occurrent* properties satisfy this condition, and this is then referred to as [QP]. The question of *which* properties are being referred to in the QE/QR debate then becomes redundant, and the onus is on [QD] to specify them.

For the first thirty years of her life, Mary is at no time allowed to see anything coloured. Apart from this systematic sensory deprivation, however, she has access to every possible teaching facility, including visual access to the world at large via a black-and-white television monitor, in order to learn about the physics and neurophysiology of colour vision. When, finally, the fully educated scientist is allowed out of the room for the first time, Jackson argues, she learns for the first time *what it is like*, qualitatively, to see colours.

> 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 1982, p 130].

As it stands, this statement is clearly open to a number of possible interpretations, some of which appear to constitute a more compelling argument than others. Before we go into the sorts of difficulties raised in response to the argument, however, it will be convenient to clarify two points

Firstly, in the experiment envisaged by Jackson the information available to Mary in her room is not just descriptive information. She has access to every possible mode of learning available to those out in the coloured world, with the single exception that she at no time has the benefit of actually seeing colour. The thrust of the argument is that until she does have access to that particular mode of learning, and thus to the experiential quality associated with actually seeing colours in the normal way, she cannot know what that experiential quality is like. Some commentators have tended to characterise Mary's epistemically restricted situation as a restriction to learning facts only by *description* [e.g., Churchland, Madell]. Jackson, however, seems to be committed merely to the thesis that physical facts should, in principle, be *describable* in the language of physical theory; not that such facts should be *conveyable* by description alone. His point is just that, apart from the information gained by direct experiential acquaintance of qualia, there is no possible *physical* information which would enable Mary to acquire those facts. If we are able to assume that she has access to all the facts about the physical domain before she leaves

her room, then, and yet fails to acquire knowledge about qualia, it follows that knowledge about qualia is not knowledge about anything physical; that qualia are non-physical. Accordingly, our discussion will proceed on the assumption that for Jackson there are two distinct modes of learning, by direct acquaintance through either actually seeing colour or by having the appropriate experiences artificially induced [e.g., by brain probes], and by any other means available in principle to the physicalist. In order to clarify the discussion which follows, then, we shall refer to the respective modes as "*knowledge by acquaintance*" and "*knowledge by demonstration*", where demonstration is understood to be any possible means of imparting physical facts to Mary other than by actually allowing her to experience qualia, either by seeing colours or by having qualia artificially induced.

The second preliminary point is somewhat confusing in Jackson's account. Thus, he says that what he is trying to establish is not that the changes Mary undergoes on her release teach her something new about herself, but about other people.

> Before she was let out, she could not have known about her experience of red, for there were no such facts to know. That physicalist and nonphysicalist alike can agree on. After she is let out, things change; and she can happily admit that she learns this; after all, some physical things will change, for instance, her brain states and their functional roles. The trouble for physicalism is that, after she sees her first ripe tomato, she will realize how impoverished her conception of the mental life of others has been all along. [Jackson 1986, p 393][31]

---

31. The idiom employed by various commentators with regard to the supposedly non-physical *properties*, or qualia, construed as universals, is remarkably diverse. Thus, what are supposedly left out by the physicalist's account are variously described as "facts", truths" and "properties". At this stage , however, we do not intend to treat this as problematic. Roughly, a fact or truth is a fact or truth about a property. If there is a fact or truth about qualia which physicalism omits, then, we shall construe this as physicalism failing to provide a full account of the constitution and character [typic classification] of these properties in PPD terms.

Since even Mary's physical states will change on her release in response to seeing colours for the first time, then, Jackson is careful not to draw any immediate inferences about any *non*-physical changes having occurred in herself. But on discovering what it is like to see red, for example, she realises that her former complete *physical* knowledge about others was incomplete; they probably have experiences of the sort she has just had. Since her physical knowledge of others was already complete, however, the experiences they probably have must be non-physical. She knew all about their physical states and dispositions before, and now acquires them herself. In addition to those, however, she acquires new experiences. Thus, from the fact that she acquires something in addition to the physical and dispositional traits already observed in others, she is able to infer that there is something non-physical associated with seeing colours, both for others and for herself.

Jackson himself represents the argument along the following lines [Jackson 1986, p 393].

Knowledge Argument 1.

1. Mary [before her release] knows everything *physical* there is to know about other people.

2. Mary [before her release] does not know *everything* there is to know about other people [because she learns something about them on her release].

Therefore,

3. There are truths about other people [and herself] which escape the physicalist story.

On this interpretation, we see that 3 can only be validly inferred from 1 and 2 on the further Leibnitzian assumption that if something is true about everything physical, but not about everything, then everything physical is not everything. Allowing ourselves that assumption as our starting point, then, we can validly infer that there is something which is not physical and hence that there are truths which escape the physicalist account [that is, knowledge

about something non-physical]. The argument can then be summarised as follows. If Mary knows every physical fact about other people, but not what it is like for them to see red, then the latter is a non-physical fact. Therefore physicalism, construed as the thesis that every fact it is possible to know is physical, must be false.

First Objection: Two Types of Physical Knowledge.

Now, it might appear that although the sort of knowledge referred to in premise 1 is quite different from the sort of knowledge referred to in premise 2, the *facts* known in the two cases are nevertheless the same. Knowing the physical facts in the first way, surely, is a matter of mastering and understanding a set of sentences or propositions, mathematical equations and so on. The physical facts about colour vision, for example, would be describable in terms of light of various wavelengths being received into the eye and thence being transduced into electrical signals in the optic nerve, ultimately finding their way into the visual cortex of the brain, and so forth. This is knowledge of the sort we would expect to be able to acquire by demonstration, as we have just outlined it. Knowing *what it is like* to see colours, on the other hand, would seem to be a matter of having representations in a prelinguistic, non-propositional medium of some kind [Madell, p 80]. But there is no reason as yet to suppose that the facts known in the two cases are distinct. Knowledge by demonstration and knowledge by acquaintance, therefore, are quite distinct forms of knowledge, but they might still be knowledge of the same facts.

Taking this line of resistance to the conclusion of the knowledge argument, the physicalist would insist that the possibility of having both forms of knowledge is perfectly compatible with physicalism. Thus, while both premises in Jackson's argument might be true, equivocation on the expression "knows" in the two premises invalidates the move to the conclusion at 3. It is quite possible for Mary to know [by demonstration] everything physical it is possible to know, and yet not know [by acquaintance] everything physical it is possible to know, about colour vision in particular. This possibility provides a way of rendering premises 1 and 2 of Jackson's argument compatible without leading to the conclusion that physicalism is incomplete. For physicalism is compatible with the

thesis that knowing [by demonstration] everything it is possible to know does not amount to or entail knowing [by acquaintance] everything it is possible to know. If this charge of equivocation on the use of "knows" is correct, then, Jackson's argument must now be modified so as to accommodate the two kinds of knowledge, as follows.

Knowledge Argument 2.

1. Mary [before her release] knows [by demonstration] everything physical there is to know about other people.

2. Mary [before her release] does not know [by acquaintance] everything there is to know about other people [because she learns something (by acquaintance) on her release].

Therefore,

3. There are truths about other people, and therefore herself, which escape the physicalist account.

In this modified form, the argument contains premises which we may assume for the moment to be true. Mary does know [by demonstration] everything physical there is to know, let us say, but at the same time she does not know [by acquaintance] what it is like to see colours. Clearly, the inference in 3 is is now invalid, even when we take into account the supplementary Leibnitzian assumption, as before, that if something is true about everything physical, but is not true about everything, then everything physical is not everything. The argument is invalid simply because premises 1 no longer implies that Mary knows *in all possible ways* everything physical there is to know. She only knows by *demonstration* everything physical there is to know. Thus, the discovery in premise 2 fails to indicate that she learns anything new about other people. It merely indicates that there is something which it is possible to know by acquaintance as well as by demonstration, but which Mary only knows by demonstration.

What happens to the argument, then, if we try to reconstruct it in such a way as to eliminate the equivocation over the use of "knows"?

Churchland finds out by setting out Jackson's argument in such a way that it is both valid and unequivocal on the use of "knows" [Churchland, 1989]. The argument then looks like this:

Knowledge Argument 3.

1. For any knowable x and for any form (f) of knowledge, if x is physical in character, then Mary knows (f) about x.

2. There is a knowable x and a form of knowledge (f) such that Mary does not know (f) about x.

Therefore,

3. There is a knowable x such that x is not physical in character.

This argument is valid. In short, it says simply that if Mary knows everything physical in all possible ways, but does not know about the experiential quality of colour vision in a certain way, then that quality cannot be physical in character. The problem is that in rendering it valid Churchland finds that he has had to take on an unwarranted premise at 1. Mary does not necessarily know in all possible ways everything physical there is to know. Specifically, she does not know [by acquaintance] what it is like to see colours. If x is a phenomenal property, say Rp, and (f) is a form of knowledge which involves direct, introspective acquaintance, then Jackson's point is that Mary does not know (f) about Rp. But since this is entirely compatible with the claim that Rp is physical, it follows that 1 might be false. Mary can only be assumed in premise 1 to know everything physical in all possible ways if knowing Rp by acquaintance is already excluded from this category. But this would be for Jackson to beg the very question he is trying to answer. By the same token, however, if Churchland simply assumes that Rp *is* physical, and can also be known by demonstration, he is also begging the same question.

Thus, the difference of opinion between Jackson and Churchland amounts to a disagreement over the truth of the claim that:

[QD]3. Qualia are occurrent irreducibly non-PPD properties.

And since the qualia-dualist claims essentially that qualia are occurrent irreducibly non-PPD properties, we can see that the disagreement at this stage amounts to no more than a restatement of the respective positions [QD] and [QP]. The issue still to be settled is whether Mary does learn a new fact on her release; a fact which she cannot learn by demonstration, or whether she merely acquires a new way of knowing a fact already known or knowable by demonstration. If we assume here that PPD facts must at least be knowable by demonstration, then, we can see that the success of Jackson's argument depends on its ability to establish that Mary learns a fact which is not knowable by demonstration. At this stage we have only his strong intuition that it is "just obvious" that she does acquire new knowledge.

Churchland's Distinction between Knowing *That* and Knowing *How*.

An attempt to accommodate Mary's extra-mural discovery within the physicalist position is sometimes developed along the following lines. What Mary acquires, on her release from the black-and-white room, is not a new *item* of knowledge at all. Rather, she merely acquires an *ability* to discriminate or even identify colours experientially. Thus, if Mary merely acquires such an ability on her release, it cannot be argued that her knowledge of the facts was incomplete before her release. Consequently, premise 1 of each of the above versions of the knowledge argument can be held true but premise 2 rejected. There just is no new item of knowledge for Mary to acquire. Knowing what it is like to see red by acquaintance, for example, becomes having the *ability* to recognise or discriminate red by visual acquaintance. But she already knew all the *facts* about colour and colour vision before she left her room. Hence, there is no reason to suppose that any facts of a non-physical or non-physico-dispositional nature are discovered on her release.

In effect, what is being claimed here is that when we employ the expression "phenomenal red", for example, or "Rp", we thereby refer to no specific physical property at all. Rather, we refer to an ability or dispositional trait of some kind which may be physically grounded. It is the dispositional type [the ability] which is to be

identified with the experience of phenomenal red, however, the physical constitution of instances of this type being unimportant. The ability hypothesis is therefore topic-neutral with regard to the type of physical state which furnishes Mary with her new-found ability. The physicalist, of course, claims at least that Mary acquires her ability in virtue of entering a token physical state of one sort or another; that when she sees red for the first time she merely acquires a neuro-physiological state which constitutes an instance of seeing red. Whether tokens which fit this description are of a particular physical type is a further question which need not be addressed at this stage.[32]

Implications of the ability hypothesis.

Both QD and QP seem to assume that on being released from her room Mary should be able to discriminate and *identify* particular colours without hesitation. Even on the ability hypothesis, however, this seems preposterous. To get this in perspective, consider the case of a sighted observer trying to teach Mary how to recognise apple green, or even more specifically, sodium yellow. Suppose that Mary were able to understand everything neurophysiological and dispositional that was said to constitute or characterise this discriminatory ability. Does it seem even remotely plausible that she might leave the room and immediately proceed to pick out a sample of sodium yellow, and distinguish it from lemon yellow, chrome yellow and many other hues which the sighted observer has learned by acquaintance to discriminate? The suggestion that she could would need vastly more substantiation than Churchland is able to provide. For although she might be expected to know all the relevant states a sighted observer gets into when seeing sodium yellow, it is quite a different matter for her to be able to

---

32. This is essentially the position offered in resistance to the inverted spectrum argument. Occurrences of Rp are to be construed as occurrences of some [topic-neutral] PPD state which engender dispositional traits of a particular type. In this particular case, the type is described as an ability to discriminate colour properties.

determine *that she is in those states*. Unlike the experienced sighted observer, she would have to know that she is in whatever physico-dispositional states are characteristic of seeing that colour in particular. If the states are neural types, it is difficult to see how she might begin. If they are dispositional types, the task confronting her is scarcely less challenging.[33]

The crucial point, however, is that this is an unnecessary burden for the physicalist. Knowing all about the physical facts associated with seeing red does not entail the *ability* to determine that one is seeing red. When Mary discovers all the physical facts about what is going on when Smith sees red, for example, she might establish that seeing red occurs whenever Smith is in a physico-dispositional state RP. Her subsequent discovery on seeing red for herself is that Rp is an additional characteristic associated with seeing red. The ability hypothesis is just that there is no such characteristic; what Mary learns on her release is how to recognise red for herself. The ability hypothesis does not entail that she should be able to do so immediately, however. Given all the physical facts, she knows that seeing red amounts to being in state RP, but this does not entail that she should *know* that she is in state RP on her release. In order to establish this fact she would need to conduct further research into the conditions obtaining at that time.

The disagreement between Jackson and the physicalist is over the question of *what she discovers on her release*. The physicalist will have to say simply that she discovers nothing at all. It is true that she acquires neural and dispositional states of certain types for the first time, and that these are uniquely associated with seeing red, but for Churchland this does not amount to the acquisition of new information. According to the ability hypothesis, there is nothing of a non-physical constitution associated with being in those states and therefore no additional facts about those states *per se* for her to learn. This has nothing to do with her ability or inability to establish epistemically which state she is in *on a particular occasion*.

---

33. See Robinson, 1993a, for a more detailed account of this sort of objection.

Reply; The "Ability" Hypothesis is Compatible with Jackson's Position.

Jackson's response to the ability hypothesis is that while it is indeed true that Mary acquires a new ability to discriminate red by acquaintance, at least over time, she does so by becoming familiar with a *phenomenal property* of a type typically associated with seeing red. To assume that she does not would again be simply to presuppose that there are no phenomenal properties.

His support for this position [Rosenthal, p 394] is that since Mary learns not only something about herself but also about others, it must be something more than the mere ability to identify colours that she acquires. She already knew all about others' abilities before she left her room. What she learns about other minds, or so it seems to him, must be more than the fact that they have certain *abilities* to identify colours. What, for example, would a Mary who doubted the existence of other minds be doubting she had learned by emerging from the room? Jackson thinks it must be something over and above mere abilities, since she could find her knowledge of others' abilities unproblematic even though doubting the existence of other minds. The new knowledge, therefore, must be knowledge in the original sense; that is, knowledge about some property or quality of experience, which may or may not occur in other minds.

Jackson's response seems inadequate, however, since it has not been made clear why Mary's newly acquired knowledge [over time] cannot be construed simply as the ability to identify colours experientially for herself; an ability she already knew others to have. Unless we were to assume that her ability *is* facilitated or at least accompanied by phenomenal properties, there is no reason to suppose that Mary will have anything about other minds to wonder about. To bring in knowledge of other minds and their contents in defence of his position simply begs the question he started out trying to answer; namely, whether or not there *are* phenomenal properties to be discovered only by experience. Unless the existence of phenomenal properties can be established by some independent demonstration, Jackson cannot argue that Mary acquires knowledge about something which it is possible to doubt that other people have.

A more modest defence of Jackson's position might seem appropriate. Thus, he might claim that it is by no means clear how Mary or anyone else could acquire an ability to identify colours experientially without there being something that it is like to do so. The point is not that there *must* be some mental property which we refer to as "what it is like to see red"' for example. Rather, the question of whether there is any such property remains unanswered by the ability hypothesis. Thus, while Mary might well learn how to identify red after she has been released, she might do so by learning *what it is like*, Rp, to see red. Knowing how to identify red by acquaintance might be having the ability to recognise a certain kind of phenomenal property. Far from offering an alternative to Jackson's thesis, then, the ability hypothesis seems to be perfectly compatible with it.

The question that must be tackled here is whether the ability to identify red experientially, which Mary is assumed to acquire, is or is not facilitated, or at least accompanied, by her experience of a phenomenal property. Since Jackson's argument assumes that there are such properties, we cannot employ the assumption that there are not in order to undermine it. At best, the question remains unanswered. For anyone who starts out with the strong intuition that Mary identifies and discriminates red by reference to what it is like to see red, the suggestion that there is no such property over and above the bare ability to discriminate red seems completely unhelpful. Neither Jackson nor the proponent of the ability hypothesis has settled the matter of whether there are phenomenal properties or not.

Mary Can Learn Rp by Demonstration, after all.

In order to sidestep these difficulties, Churchland comes up with the rather surprising assertion that Rp is a knowable item after all, but that Mary *is* able to discover Rp even before she leaves the room. At least, it is possible in principle that she should be able, on the basis of her complete knowledge of neurophysiology, to *imagine or work out* Rp. To suppose otherwise, he argues, would be to presuppose without sufficient evidence that phenomenal properties are beyond the reach of physical demonstration. But his account appears to equivocate between the construal of qualia as

neurophysiological items and as something additional to be worked out or imagined.

> In particular, suppose that Mary has learned to conceptualise her inner life, even in introspection, in terms of the completed neuroscience we are to imagine. So she does not identify her visual sensations crudely as "a sensation of black", "a sensation of gray", or "a sensation of white"; rather she identifies them more revealingly as various spiking frequencies in the nth layer of the occipital cortex (or whatever). If Mary has the relevant neuroscientific concepts for the sensational states at issue (viz., sensations-of-red), but has never yet been in those states, she may well be able to imagine being in the relevant cortical state, and *imagine* it with substantial success, even in advance of receiving external stimuli that would *actually produce it.*" [Churchland, 1985 pp 25-6] (My emphasis).

It seems that only the first interpretation is available to Churchland, since if qualia are physical properties they will not need to be worked out, given a complete neurophysiological knowledge of seeing colours. Jackson's qualia discourse would therefore be construed by QP simply as [*ontologically* redundant] discourse about already accepted physical members of [S]. But if Mary knows all the physical facts, it follows that there is nothing further for her to imagine. She knows all about the states in question, but has not yet been in them. There is nothing that it is like, however, to be in those states, and therefore nothing further for Mary to work out. QP might concede, nevertheless, that such discourse is not *epistemically* redundant. Thus, he might concede that in certain epistemic situations Smith can determine that he has a headache even if he does not *know* that a headache is physically constituted. In that case the type [headache] is epistemically topic-neutral with respect to the physical state which constitutes it. So he might concede that Mary can know that she has a sensation-of-red even before she learns what physical state it amounts to. The crucial claim for the physicalist is just the ontological one that there are no additional properties which the current physicalist account of experience leaves out.

Dennett might also be interpreted as subscribing to the view that there is nothing "in principle" to prevent Mary from learning what

it is like to see colours before she leaves her room. The crucial point for him is that if we are to imagine her doing so she must be imagined as knowing everything physical, not just "lots and lots" [1991, p 399] about the physical world.

> Simply imagining that Mary knows a lot, and leaving it at that, is not a good way to figure out the implications of the hypothesis of her having "*all* the physical information". [Dennett, 1991, p 400]

It is not entirely clear, again, whether Dennett is suggesting here that the ontic commitments of a future neuroscience will be more comprehensive than those of the current physics. It appears that he is "living on the edge" [Dennett, 1993] already occupied by Churchland. If he insists just that the experiential character of phenomenal properties could be "figured out", given all the currently acknowledged physical facts, he must presumably be implying that they are not identical with any properties already counted as physical. On the other hand, there are times when he seems to believe that they are. Thus, he imagines the first coloured object the omniscient scientist is confronted with as being a blue banana. She is not fooled, however.

> Mary took one look at it and said "Hey! You tried to trick me! Bananas are yellow, but this one is blue!". [Dennett p 399]

Her explanation of how she was able to perform this feat was as follows.

> "You have to remember that I know everything - absolutely everything - that could ever be known about the causes and effects of color vision. So of course before you brought the banana in, I had already written down, in exquisite detail, exactly what physical impression a yellow object or a blue object ... would make on my nervous system. .....I realize it is hard for you to imagine that I could know so much about my reactive dispositions that the way blue affected me came as no surprise. ..... Its hard for anyone to imagine the consequences of someone knowing absolutely everything physical about anything!" [Dennett p 400]

This might be interpreted as the claim that all the facts about colour vision will ultimately be accepted as being physical facts

which we already accept, rather than as being additional facts to be worked out. Churchland's position on this point might be similarly interpreted. There is no reason to suppose, he says, that some future "completed neuroscience" will, like our present neuroscience, be incapable of conveying the experiential character of Rp to Mary purely by demonstration. But whether the completed neuroscience will show Rp to be a property already recognised by physics or merely enable Mary to imagine or work out Rp as an additional property is unclear.

Let us examine the two positions between which Churchland and Dennett seem to be wavering in more detail, but now paying special attention to our requirement that for reductive physicalism to be true, a complete account of qualia must be possible by recourse the *already accepted* members of [S].

1. Qualia-Reductivism.

If this is the thesis being expounded, the claim is that there are no occurrent qualia in addition to the PPD states and properties already acknowledged as belonging to [S]. Thus, qualia might be neural properties per se, or experiencing qualia might be construed as exhibiting the [neurally grounded] ability or disposition to discriminate colours. Some philosophers express the thesis in rather different terms, however, by suggesting that a successful programme of reduction of the phenomenal to the physical would have to leave the former *logically inferrable* from the latter:

> The reductionist looks at his analysans and sees that it entails that which he is reducing: the eliminativist considers his preferred theory and sees that it does everything legitimate that was done by the eliminated theory, but that it does not entail it [i.e., it is redundant]. [Robinson, 1994, p181] (My parentheses)

But Robinson's portrayal of reduction here must be approached with caution in the present context. Thus, we saw in chapter III that Smith can determine in Eq that he has an item of the type [headache] even when he is unable to determine that it is a physical item. But this is made possible by the *epistemic* topic-neutrality of the type [headache] with respect to any physical type which the headache

might belong to [as Robinson also acknowledges (op.cit., p 182)]. Robinson's explanation of why the inferential route must obtain is that:

> Scientific knowledge of the brain is topic specific and is more detailed than topic neutral experiential knowledge of the brain; the scientific knowledge, therefore, *includes* everything the experiential knowledge contains. [op.cit., p182].

So while it is possible for someone to discern in Eq [topic-neutrally] that what he is experiencing is a headache Hp, he might do so without knowing all the physical facts about his headache. On the other hand, if he determines by physiological investigation that he is in a state of the type HP, physicalism ordains that his episode of introspecting his headache just is an occurrence of HP. All the facts about his introspective episode are physical facts about members of [S].

We must be careful here to interpret Robinson's appeal to entailment in the correct way. Thus, for the physicalist, the complete physical facts about qualia; what it amounts to for something *to be* a quale, do *not* entail any facts about what is occurring *at any particular time*. For we saw that a Mary who has a complete grasp of physiology *per se*, including an understanding of what is going on when someone sees red, should not be expected to recognise particular colours immediately on emerging from her room; it is a matter for further investigation to discover which PPD state she might be in on that occasion, and thence which colour she is seeing. Similarly, then, someone who has a complete grasp of physiological theory should not be expected to infer from that knowledge what is going on for Smith introspectively; further investigation will be required. In both cases, a complete grasp of the theory does not entail a knowledge of what is occurring at any particular time. The relevant distinction here is between having a theoretical understanding of the nature of qualia and empirical knowledge about whether such qualia, as explained by that theory, are occurring at any particular time.

What Robinson is referring to, then, is the relation of entailment between the qualia-facts as expounded by physical theory and the qualia facts which are discernible in introspection. Whatever is occurring when Smith discerns a quale in introspection, all the

facts about that occurrence must be physical facts. Hence, if we understand the physical facts about qualia we understand everything about what is occurring for Smith as discerned in introspection. But it requires a posteriori investigation to discover what *is* going on, even in terms of physical theory. In terms of properties, construed as universals, there are no occurrent properties for the physicalist which cannot be fully accounted for within the terms of the physical theory. Hence, if Smith really does discern specific properties in introspection, those properties, and indeed the introspective episode itself, are fully accountable in physical terms. But the relation here is not one of entailment, but of *identity*. So we might be more accurate if we say that the analysans does not *entail* that which it is reducing; rather it just *is* that which it is reducing.

The question we need to ask, then, is *which* theoretical facts Robinson thinks are left out by the physicalist. Thus, suppose that the physicalist adopts the ability hypothesis, and claims that recognising Rp amounts just to determining that a physical state token is of the type [Rp]. Robinson's point is then that if the physicalist were to cite RP as a token brain state which can be classified, topic-neutrally, as a type [Rp] item in introspection, this would entail that the type [Rp] is fully accountable in physical terms; that Rp is just a physical property already contained in [S]. And this amounts to our stipulation in the introduction that reductive physicalism must provide a full account of all occurrent *types*, where an item is taken to be of a particular type [Rp] if it has the property Rp. The question confronting the dualist, then, is which particular occurrent properties are not included in [S].

The major difficulty here is that the properties discerned in introspection are discerned topic-neutrally. Hence, from the first-person perspective, even the physicalist cannot be expected to determine which physical properties these are without conducting an a posteriori investigation; that is, by investigating what is going on in the third-person perspective. But this amounts just to acquiring further mastery over the current physiology of introspected properties. Thus, if our current mastery of physiology is less than comprehensive, even within the terms of current theory, it might still turn out, as an a posteriori discovery, that introspected properties are just physiological properties. In order

to rule out this possibility, then, some fact about introspected properties must be cited which precludes their being physiological properties [members of [S] even in principle. For we might fairly assume that none of us yet knows *all* the physiological facts about introspection and its properties, even as couched in current physiological theory.

In the absence of a complete knowledge of all the physiological facts about introspection, Robinson might resort to the further claim that there is *something that it is like*, Rp, to see red. Even though the current physiology should be capable of accounting for all of the physiological facts about introspected properties, then, Robinson's claim must be that it would still not contain all the information about Rp itself, as an experiential character.

2. Qualia Incorporativism.

In this case, what Dennett and Churchland would have to argue is that although phenomenal properties cannot be reduced to physico-dispositional constitution and character within the context of our current physiological theory [i.e., they are not members of [S]], a *completed* physical science might nevertheless incorporate them as additional properties. Incorporativism can only be compatible with physicalism if it is construed as the thesis that qualia are, after all, to be included in the catalogue of physico-dispositional items acknowledged by the completed science. But on that interpretation it is even more difficult to see how the physicalist's claim can be ruled out in principle. For we have no indication as yet that the completed science conceived by the physicalist will not be ontically committed to qualia. What is needed here is a more far-reaching fact about qualia which precludes the possibility of their inclusion in any *possible* neuroscience. And this immediately invites the question of what conditions any neuroscience would have to comply with.

As it stands, then, premise 2 of the knowledge argument, that Mary is unable to acquire a knowledge of qualia given all the physical facts, remains unsubstantiated. Our particular interest at present is in the *reductive* version of physicalism already explained. For if we can find no convincing reason for precluding qualia in principle

from the current physical ontology, then a fortiori the possibility of incorporative physicalism being true cannot be ruled out either.

Jackson's Position

The argument presented by Jackson can now be summarised in the following way. Firstly, premise 1 states that even in the black-and-white room Mary can know all the physical facts about what happens to other people when they see red. Secondly, however, premise 2 claims that when she emerges from her room and sees red for the first time, she is able to determine that what she experiences, Rp, is additional to all the facts she already knows about other people. It then follows from the fact that her physical knowledge of what happens to others when they see red is complete that Rp is an item of non-physical knowledge.

In the light of the foregoing discussion, we can now see that Jackson's disagreement with the physicalist can be simplified. Firstly, the debate as to whether there is *something that it is like* to see red is no longer relevant. Even the physicalist can concede that there is something that it is like, quale Rp, *insofar as* the expression "quale Rp" can be construed as a reference to a physical state which is associated with seeing red. He has no intelligible way of being more specific than this. Indeed, David Lewis adopts this position explicitly. Thus, confronted with the problem of having to admit that he knows what it is like to taste Vegemite, he incorporates this undeniable fact into his physicalistic thesis by claiming that:

> There is a state of knowing what its like, sure enough. And Vegemite has a special power to produce that state. But phenomenal information and its special subject matter do not exist. [Rosenthal, p 234] [Lewis reassures us that Vegemite is a celebrated yeast-based condiment; presumably, he is anxious not to mislead his readers into experimenting at home with any form of explosive material].

Secondly, then, it is just the *completeness* of the physical account which is the point of contention. Thus, while Jackson might refer specifically to *qualia* as non-physical properties, the physicalist will construe this as the simpler claim that there are non-physical

properties, and thence infer that Jackson is mistaken. The onus is therefore squarely on Jackson to come up with an intelligible criterion for the physical; he must explain what it would amount to for an item to be physical, and then cite an occurrent item which fails to satisfy that criterion. Clearly, what is needed in support of the *knowledge argument* is some sufficient reason for positing an Rp which cannot, even in principle, be conveyed to Mary by all the available physical evidence. More specifically, he must employ a distinction between two different types of property; physical properties, which for whatever reason must all be conveyable to Mary by demonstration, and phenomenal properties, which must not. Only if he has shown that phenomenal properties are of such a type as not to be conveyable by physical demonstration will he be entitled to draw the specific conclusion he needs about Rp; that since no possible physical demonstration can convey Rp to Mary, Rp must be non-physical.

Suppose that Mary has a companion, Smith, in her room and that Smith is looking out onto the coloured vista through a private window. His job is to act as Mary's guinea-pig. Thus, when he sees something red he not only does his utmost to explain what it is like to see red and displays all the red-appropriate dispositions, but allows himself to be subjected to any scientific studies Mary might care to conduct. Her evidence for what Rp is like experientially for Smith then includes verbal descriptions, behavioural responses and the results of any possible neural scanning or probing experiments. Jackson must concede that whatever facts are available for Mary in this situation, she has access, at least in principle, to all the physical facts. She has access to all of Smith's non-descriptive reactions and neural states, in addition to his description of Rp couched in terms of current neuroscience. The claim must then be that even given all this evidence there is something about what happens to Smith when he sees red which she has yet to learn. However plausible it might seem that her knowledge will be incomplete, however, we have as yet no sufficient reason for insisting that it is. As we saw in chapter II, [QP] entails that all of the available evidence from the third-person point of view leaves the positing of *non*-physical qualia unjustified *by* the physical facts. But since QP is claiming here that the available facts are all *physical*, the question of whether non-physical qualia can be justifiably posited on the basis of *non*-physical evidence, and even

known, remains unanswered. We have, as yet, no way of denying the possibility outright without presupposing that physicalism is true.

If we refer to Smith in this situation as offering an *optimal physical demonstration* [OPD] of seeing red, we need a good reason for insisting that there is an experiential character [which Jackson might refer to as the quale Rp], which Mary is unable to learn from the OPD offered by Smith.

Public Physical Facts and Private Mental Facts.

The argument we are trying to produce in support of Jackson is now required to take the following form. Our reference to *properties* here is intended as a reference to properties as universals. *Conveying* a property amounts to successfully imparting, interpersonally, a complete understanding of what a particular property amounts to.

Knowledge Argument 4.

1. All physical properties are conveyable by OPD.

2. Phenomenal property Rp is not conveyable by OPD.

Therefore,

3. Rp is not a physical property.

At the same time, we need to leave open the possibility that Rp is a *mental* property. Thus, we might add the following premise and conclusion.

4. Not all mental properties are conveyable by OPD [where OPD includes even mental description].

Therefore,

5. Rp might still be a mental property.

If premise 1 is taken as an indication of what a property would have to be in order to be counted as physical, this argument is directed against *all* forms of physicalistic account. Thus, if premise 2 were true, it could be inferred that even Churchland's future "completed neuroscience" will not incorporate Rp. If it is successful in refuting all such accounts, then in the absence of any further categories in which to fit phenomenal properties we would be in a position to conclude that they must therefore be mental. This is a valid argument, clearly, but what about the premises?

Premise 1 can be construed in this way because Jackson has set up his experiment in such a way that Mary has all the evidence available by OPD. Thus, he claims that even before Mary has left the room, she has access to all the physical facts about other people. We are entitled to assume, then, that whatever demonstration is available to her is sufficient to convey those facts. Hence, if an OPD is construed just as a demonstration of all the physical facts, Mary has access to an OPD. It is only by discovering facts outside the room which she does not already have about other people that she is able to draw the inference that she has acquired knowledge of non-physical facts. Even outside the room, however, Mary is only able to *wonder* whether other people have the experience she has just had; she is able to wonder whether her knowledge obtained by OPD about what happens to other people when they see red is incomplete in that they, *also*, have such experiences. Hence, if they do have such experiences, the fact is not conveyable interpersonally. This can be misleading, however. Jackson's premise 2 is not that such facts are not conveyable interpersonally *per se*; rather, it is that such facts are not conveyable by *OPD*.

Hence, the significance of premise 1 for Jackson would depend on there also being some way of defending premise 2. He must be able to support his claim that there are facts about seeing red which cannot be obtained by OPD but are knowable by direct introspective acquaintance. If the OPD has conveyed all the physical facts, it will then follow that Rp is a non-physical property. Without having acquired a complete mastery over current neuroscience, however, it appears that he can only support premise 2 by claiming that it would be necessary to experience Rp in introspection in order to know what it is like. And this is substantially the claim that Rp is knowable

only by direct experience in the first-person; that it is not conveyable interpersonally *per se*.

If we take premise 1 as our starting point, then, attention must be focussed on premise 2 As we observed earlier, no prima facie case has yet been established for rejecting the possibility that if Mary really had *all* of the physical information [OPD], not just "lots and lots" of it, she would know everything there is to know about seeing colours. Churchland and Dennett make it clear that they subscribe to this possibility. There could be some improved neuroscience in the context of which all the facts are, after all, fully conveyable by OPD. In terms of our present discussion, we can say that even within the context of current neuroscience it might still turn out that Rp *is* conveyable by OPD.

Justification for Premise 1.

The support we are proposing for premise 1 might be that all physical facts must be conveyable by OPD because the physical world or domain is necessarily inhabited entirely by items of a third-person nature; that is, items which are publicly demonstrable; accessible to objective, or consensus, scrutiny. This is how we tentatively construed the physical in the introduction. We are not claiming at this stage to have demonstrated that this is an appropriate characterisation of physical items, of course. The point here is simply to explore the consequences of such an account for Jackson's argument. The supporting argument would run along the following lines.

[A]. Every physical property is publicly demonstrable.

Hence,

[B]. Every physical property is conveyable by OPD.

In turn, we could then use the conclusion at [B] to verify premise 1 in the original knowledge argument. Thus, after rewording, premise 1 becomes:

1. Mary knows every physical fact about other people [because she has access to the OPD].

This is at least one way in which premise 1 of the present argument, in turn offering support for premise 1 of Jackson's argument, might be supported.

Justification for premise 2.

Premise 2 of Jackson's argument can be viewed in a similar light. Thus, his claim that Mary would be unable to learn what it is like to see red purely by OPD must be founded on some observation regarding the incapacity of OPDs to convey the experiential quality Rp interpersonally. The premise might then amount to:

[B'] Rp is not conveyable by OPD.

We would then hope to use the premise [B'] in support of Jackson's premise 2 thus:

2. Mary does not know every fact about other people [because she does not know Rp].

But we can see now that if we take the two premises [B] and [B'] together it is possible to construct an alternative to the knowledge argument. Thus, if:

[B]. Every physical property is conveyable by OPD.

And:

[B'] Rp is not conveyable by OPD.

Then:

[C] Rp is not a physical property [which is Jackson's conclusion].

In other words, Jackson's conclusion follows directly from [B] and [B']. By assuming the truth of [B] and [B'], then, we render

Jackson's argument redundant. If we do not assume the truth of [B] or [B'], however, it is difficult to see how else Jackson could justify his premises 1 and 2, other than by simply appealing to the fact that he regards them as being "just obviously" true. As we have seen, such an appeal cannot be taken seriously. Our response to the present suggestion must therefore be that it leads to the conclusion that whether phenomenal properties are physical or non-physical is not indicated by the outcome of Jackson's experiment at all. Only if all physical facts are conveyable by OPD is Jackson entitled to premise [B] and thence to premise 1. But in that case he is not entitled to [B'], since [B'] remains only an a posteriori possibility. But if he is not entitled to [B'] it follows that he is not entitled to premise 2. Premise 2 can only be true if Rp is, indeed, not conveyable by OPD.

Nagel's Argument.

Thomas Nagel [1974] produces an anti-physicalist argument which is based on almost exactly the two premises, [B] and [B']. His line of argument runs as follows.

1. It is possible to know all the physical facts about a creature without taking the point of view of that creature [i.e., the physical facts are publicly conveyable].

Note that Nagel's choice of a non-human creature is appropriate here in that it indicates that conveyability of an experiential character to someone who has *yet to experience that character* is being considered. Thus, while it might be plausible to suppose that two humans have similar qualitative experiences when seeing red, and thence that this fact is demonstrable by analogy [Russell, 1948; in Rosenthal, pp 89 - 91], it is much less plausible that qualitative experience is conveyable to someone like Mary who has never had that experience. It is therefore at least equally implausible that "what it is like to be a bat" should be conveyable to a human being. Hence:

2. It is impossible to know what it is like to have the sensations of another creature unless one is capable of

taking the point of view of that creature [i.e., Rp-Bat, for example, is not publicly conveyable to humans].

Therefore,

3. What it is like to have the sensations of another creature is not a physical fact.

Objections to Nagel's Argument.

The standard objection to this form of argument, according to Hill, is that any attempt to tighten it up will ultimately commit one of two possible errors. It will either use Leibnitz's law in a fallacious way, by applying it to the intentional concept "to know about", or it begs the question of whether in fact what it is like to have sensations is identical with some physical facts. Let us therefore consider each of Hill's predictions in turn [Hill; pp 88 - 90]

On one interpretation, "to know about" can be construed as something like "to have an adequate concept of". On this interpretation, it is clear that the conclusion at 3 can only be reached fallaciously. It is rather like arguing as follows.

1. Mary knows about [has an adequate concept of] heat.

2. Mary does not know about [have an adequate concept of] molecular kinetic energy.

Therefore, by Leibnitz's law,

3. Heat is not molecular kinetic energy.

The conclusion here is false, but the premises true, and the deductive error occurs, according to Hill, when we apply Leibnitz's law to expressions which are too closely related to propositional attitudes. Thus, whether Mary "knows about" molecular kinetic energy depends, on one interpretation, on whether she is familiar with the scientific theory that states the identity of heat with molecular kinetic energy. The fact that there are certain propositions

containing "knows about" which are true of heat but not of molecular kinetic energy is compatible with the two being identical. That, at least, is Hill's analysis of Nagel's argument.

The fact is, of course, that Hill's first point can be easily circumvented. On the assumption that heat is molecular kinetic energy, "knowing about heat" can be construed *topic-specifically* as knowing about molecular kinetic energy, even if the two are not known to be identical. On that interpretation, it is not true that Mary could have an adequate concept of one but not the other. If heat *is* molecular kinetic energy she clearly does not have an adequate concept of heat unless she has an adequate concept of molecular kinetic energy. The question of whether she knows the two to be identical or even has any grasp of molecular theory is logically independent of these considerations. Hence, Leibnitz's law need not be flouted in the way suggested by Hill. But as Hill suggests, this does indeed appear to lead to his second problem. If knowing about heat amounts to knowing about molecular kinetic energy, premise 2 is incompatible with premise 1.

Nagel's argument is formulated in terms of what is or is not *knowable*, rather than what is actually *known*, by Mary or anyone else. Expressed in either way, however, it is apparent that knowing about heat entails knowing about molecular kinetic energy, in the non-propositional, topic-specific sense just indicated, whether or not the two are known to be identical. Hence, the contingent shortcomings in Mary's physical knowledge at present, specifically, the fact that she does not know the two to be identical, do not entail that she does not know about molecular kinetic energy. What she might still not know, however, is *that* the two are identical. Once that has been accepted, we cannot know that premise 2 in Nagel's argument is true until we have decided that what it is like to have the sensations of another creature [Rp-Bat, for example] is not a physical fact. But that is just what we are trying to prove! The point is that if Rp-Bat were identical with some physical or functional fact then it would not be impossible to know the Rp-bat without taking the point of view of the bat. In order to know the Rp of another creature we would simply have to know the physical fact with which Rp is identical. Hence, this possibility could only be ruled out once all the physical facts were known. Since we have

already made this observation in relation to Jackson's version of the argument, we need have no difficulty here in agreeing with Hill.

Cynthia Macdonald makes the same point, but this time using a more general psychophysical identity claim as her example. She represents Nagel's argument as follows. [MacDonald, pp 20- 21]

1. Physical types are knowable from infinitely many points of view.

2. Sensation types are knowable subjectively only.

Therefore,

3. Sensation types are not physical types.

Premise 2. then begs the question in the following way.

> Against the [second premise] it might be objected that the most that Nagel is entitled to is the claim that sensations are known subjectively only, not that they are knowable subjectively only. Suppose that pain is indeed C-fibre stimulation [an issue that should be open at this stage of the argument]. Then, being a physical type, it is [if premise 2. is true] knowable, if not known, from other points of view." [MacDonald p 21]

Macdonald then goes on to conclude that the word "only" must be removed from premise 2 to avoid begging the question. We have to allow the possibility at this stage that sensation types are physical types and as such also knowable from other points of view. Premise 1 is dealt with similarly. To assume that physical types are not knowable subjectively is to presuppose that they are not identical with sensation types. The premise must therefore be so interpreted as to allow the possibility that physical types can be known subjectively too.

Our response to MacDonald must be that the known/knowable issue is irrelevant to the charge of question-begging. Even if "known" is used in both premises, the question is begged in exactly the same way unless "known" is construed propositionally. Thus, in premise 2, we cannot even claim that sensation types are *known* subjectively

only without presupposing that they are not identical with physical types. For if they were so identical, we would already know them just by knowing the physical types. To suppose otherwise would be to reintroduce the propositional difficulties already cleared up. What we would not know, possibly, is *that* those known physical types are in fact sensation types. The important point is that the use or covert assumption of the word "only" in either premise is the source of the problem whether "known" or "knowable" is used. Thus, consider how the argument would look in each case.

Argument using "Known".

1'. Physical types are known from infinitely many points of view [and also subjectively].

2'. Sensation types are known subjectively only.

Therefore,

3'. Sensation types are not physical types.

Argument using "Knowable"

1". Physical types are knowable from infinitely many points of view [and also subjectively].

2". Sensation types are knowable subjectively only.

Therefore,

3'. Sensation types are not physical types.

Clearly, premises 2' and 2" presuppose the truth of 3' in exactly the same way.

The crucial error in Macdonald's response can be most easily explained by reference to the earlier discussion of one of Churchland's objections to Jackson. Churchland, it will be recalled, considered that what Mary lacked in her room was merely knowledge of

what it is like to see red by acquaintance. She already had that knowledge by OPD. Our objection to this approach was that, given the assumption that there is something about knowledge by acquaintance which is not also a feature of knowledge by OPD, that is, the experiential property Rp, then Churchland's analysis simply omits Rp from the discussion. The proposal that knowledge by acquaintance and knowledge by OPD can be about the same physical [or PPD] property just fails to tell us anything about the specifically experiential property we want to explain. Hence, in order to rebut Churchland's objection, Jackson needs to establish that Rp is known, or knowable, *only* by acquaintance. Simply to assume that this is the case is to beg the question of how informative a completed neuroscience [even the current one] might be.

The Distinction between Nagel's Argument and Jackson's Argument.

A crucial difference between Nagel's argument and Jackson's argument is supposed to be that Jackson alone allows the possibility that qualia might be knowable in the third-person. All he insists on is that that they cannot be conveyed to Mary, in the third-person, by any *physical* information. Hence, they cannot be included in the physical ontology. We have been construing Jackson's argument as being based on the tacit premise that all physical items, properties, etc. are at least *conveyable by OPD*, irrespective of whether they are interpersonally conveyable by any other means. For the experimental situation involving Mary to be relevant, however, this has to be recast as the premise that all physical items are conveyable interpersonally by OPD, *even to someone who has never been acquainted with them.* The argument then proceeds along the following lines. All physical items are conveyable by OPD, but qualia are not. Hence, qualia are not physical items. Thus, Jackson's argument appears not to depend on Nagel's assumption that qualia are knowable only subjectively. Even if they are knowable *objectively*, they are not conveyable by OPD to someone who has never experienced them.

In order to support his second premise without recourse to Nagel's assumption, then, Jackson would have to find some other support for his more specific claim that qualia are not conveyable by OPD. In other words, he must be able to cite some other fact about qualia

which entails that they cannot be conveyed by OPD even if there might be some other possible means of conveyance. In common with most advocates of Jackson's approach, we can assume for the argument that this fact is just that qualia are "raw feels", and at this stage we need not worry about what a "raw feel" might amount to. But this invites the question of how he *knows* that raw feels are not conveyable by OPD. For unless he can respond convincingly to this question, it remains possible that raw feels are, by his own criterion, just physical properties. He would then have to cite some *further* property of raw feels even to preclude their reduction to already acknowledged members of [S].

If he replies that raw feels are not conveyable by OPD because no physical description has anything to say about them, he is then vulnerable to Hill's first objection. Thus, it might be the case that an acknowledged physical property RP [which is conveyable by OPD] is identical with a raw feel, even if Jackson does not know that the proposition that a raw feel is RP is true. The "raw feel" which Jackson wants to know about might be, in fact, just an unwitting topic-neutral reference to RP; a fact to be discovered by further a posteriori investigation. Once the propositional difficulty has been cleared up, then, there is a second problem. For if Jackson knows RP, it might turn out on a posteriori investigation [i.e., on gaining further information within the context of current physiology] that the raw feel *is* RP. So he is obliged to counter this objection by offering further evidence that it is not.

The evidence resorted to is typically that raw feels are non-physical because:

1. It is essential to certain types of mental phenomena that they feel in characteristic ways to their subjects [MacDonald, p 27].

Raw feels are then to be distinguished from physical properties on the grounds that:

2. It is essential to raw feels that they feel in character-istic ways to their subjects.

But:

3. It is not essential to any physical phenomenon that it feel in any characteristic way to its subject.

And from these considerations it is hoped that raw feels will be successfully distinguished from all physical phenomena.[34] But there is more work to do, for it is not yet clear that premise 3 is true. Clearly, it is not acceptable to justify 3 by resorting to the previous point that physical description has nothing to say about raw feels, or that it is possible to know all the physical facts without knowing raw feels, since then the argument would be circular. What other evidence might be cited, then?

MacDonald supports Jackson by arguing that:

> Experiential, phenomenal, properties evidently are not [as physical properties are] capable of possession by subjects that lack consciousness. [p27].

If raw feels are phenomena which are essentially possessed by conscious subjects, however, Jackson's argument might now be recast in the following way.

1. No physical property is essentially possessed by a conscious subject.

2. Qualia are properties essentially possessed by a conscious subject.

Therefore,

3. Qualia are non-physical properties.

---

34. Notice that if this approach were successful, it would effectively preclude the possibility even of non-reductively annexing qualia on to the agreed set [S] as *additional* physical properties. For if it were true that *all* physical properties, but not raw feels, essentially have the subjective property cited, it would follow that raw feels, even in principle, cannot turn out to be physical properties at all.

But this again gives Jackson the result he needs without reference to the knowledge argument. If the conclusion is sustainable in the way outlined, the knowledge argument is redundant. On the other hand, there is no immediately obvious way of supporting the above premises. Premise 1 might be taken as an indication of what is meant by a "physical property", but then premise 2 needs support. Simply to assume that it is true would be to beg the question as to whether qualia might be physical items which *can* be, but are not essentially, possessed by a conscious subject.

In order to avoid begging the question, then, Jackson will have to approach his premise 2 in some other way. He must establish that raw feels are not conveyable by OPD without presupposing that raw feels are essentially possessed by a conscious subject. Suppose, then, that he argues that raw feels are not conveyable by OPD because:

1. Raw feels are not interpersonally conveyable by any means whatever and are therefore not conveyable by OPD.

But this is just Nagel's premise 2; that qualia are knowable subjectively only. In that case, premise 1 of the knowledge argument no longer needs to appeal to OPD at all. The argument remains valid even if premise 1 is just that all physical phenomena are intersubjectively conveyable. And this renders Jackson's argument identical with Nagel's.

It seems that Jackson has three possible options open to him. Firstly, he might insist that qualia are knowable only subjectively, and concede that his argument is just a restatement of Nagel's argument. As we have seen, this leaves open the question as to how his premises might be substantiated. How, in particular, might he establish that items which are knowable subjectively are not just items which are already known objectively. Secondly, he can abandon the knowledge argument altogether and rely on the assertion that qualia, but not physical items, are *essentially* possessed by conscious subjects. As we shall see in the next chapter, either of these options might be supported by the modal argument offered by Kripke, if that argument succeeds. The third possibility is that he can try to find some other justification for his second premise, that even if conveyable by some means or other, qualia are not conveyable by OPD.

The only third possibility I can see is the one advanced earlier by Robinson; that if qualia were physical items it would be possible to infer everything about them from the already acknowledged physical facts. But since, as we assumed, Robinson does not yet *have* all the physical facts, he cannot yet know that Mary would have anything to learn when she leaves her room. It seems that he can only substantiate his claim at present if he can find some other fact about qualia which precludes them in principle from physicalistic reduction or annexation. Until he is able to do so, the second premise of the knowledge argument remains unsupported.

Robinson finds this objection tedious [1991, p162]. He patiently explains that:

> the notion of *knowing everything* is merely an aid to easy exposition of the argument. It can be expressed without it. The crucial idea behind the argument is that no possible knowledge of a physical sort would constitute or entail knowledge of the subjective dimension. [1993, pp 162-3].

In terms of his example of a deaf scientist, DS, his premise 2 of the knowledge argument becomes his premise 18, that:

> Whichever set [of facts in principle expressible in the vocabulary of physical science] that DS knows, he, unlike those who can hear, does not know the phenomenal nature of sound. [p 163] (my parentheses).

But without having access to all such physical facts, Robinson is pinning his claim on the fact that the qualia in question are items in the "subjective dimension". If he is to avoid appeal to Nagel's question-begging premise that subjectively knowable qualia are not also objectively knowable items, then, he needs some other way of substantiating his premise 2 and thus rendering his argument sound. But he openly admits to having no other way. As far as any dispositional analysis is concerned, his position is that:

> The claim that what DS lacks is more than a mere ability is not something that the argument proper proves, rather it *presupposes* it. .... a clear-headed behaviourist, functionalist, or causal theorist would always have realized that he was obliged to treat experience as no more than a dispositional state, and not as a state characterized by

> knowledge of *what it is like* in any stronger sense than knowing how. [1993, pp 182-3].

If the knowledge argument *presupposes* that the dispositional account is false because it leaves out any account of qualia, then, what about Churchland's neurophysiological account of qualia? Clearly, since dispositions must, for the physicalist, be at least neurophysiologically grounded, a similar stance with regard to the latter might be expected. And, indeed, Robinson's stance with regard to the latter *is* similar. Thus, he says that:

> What makes certain kinds of neural representations, and not others, constitute experiences? Churchland's account of *qualia* in terms of 'state spaces' does not seem to touch this problem. It is difficult to see how any of his neurology could be relevant unless there were a covert assumption that having one sort of representation *felt different* from having another. But this, of course, is what the physicalist is trying to analyse. [1993, p 168].

Again, no evidence is provided to show that what is difficult for Robinson to see must be impossible. Having identified Churchland's need to show how any neural state might feel in any way at all, he offers no evidence that it might not.

In response to the dispositional account, then, Robinson has to *presuppose* that felt qualities are not captured by any dispositional characterisation, and against the neural account he again has to presuppose that, not only does he find it difficult to see how any neural type might be a felt quality, but felt qualities are *not* neural types. He is under no illusion regarding the presupposition involved in each case, and nor is he under any illusion that he has anything more than intuition on which to base it. But this again leaves the knowledge argument hanging on the assumption that subjectively knowable qualia are not any of the items known to be conveyable by OPD. In order to turn this assumption into a demonstrated fact, and thence infer that qualia are not PPD items, further evidence is needed. If a complete grasp of all of the physiological facts, within the context of current physiology or even the future neuroscience envisaged by Churchland, are unavailable at present, some other means is again needed to turn the unsubstantiated assumption into a demonstrated fact. Once again

then, it appears that we can do no better than appeal to Kripke's demonstration that no subjectively knowable qualia *can* be identical with any objectively knowable physical properties.[35]

Jackson's Discovery about Other People.

Perhaps the clearest way of drawing the essential elements of the discussion together and focussing on the crucial problems facing Jackson's argument is by reference to another thought experiment. In order to keep the example as uncluttered as possible we can assume that if the ontic commitments of current science are contained in the set [S] of physical items, the future completed neuroscience envisaged by Churchland shares [S]. Thus, for the physicalist, there are no additional or revised ontic commitments to be allowed for. Further, we can assume that premise 1 of Jackson's argument is true; that an OPD of Smith's PPD states and characteristics is available

---

35. A further option open to the dualist here might be that what we are to *count* as "the physical" precludes any items which are knowable subjectively. Thus, qualia would then be non-physical just in virtue of being knowable subjectively. The problem with this suggestion, however, is that even paradigmatically physical items are knowable subjectively. Thus, even for the physicalist, Smith in Eq can determine that he has a headache even if the headache is a physical item. What he might not know is *that* it is a physical item. To preclude his headache from membership of [S] on the grounds that he can know that he has it subjectively, or in introspection, would amount to presupposing that his headache is not physical. For we are surely not willing to employ a criterion for the physical which is based on such epistemic considerations. Even a purely physical robot might be in such an epistemic situation with regard to its own states. The alternative of requiring that only items which can be known [topic-specifically] in introspection *to be* physical items are to be counted as physical seems equally unacceptable. For even the robot might be incapable of determining topic-specifically that the states it discerns in introspection are physical states, but we would not then infer that they are *not* physical states.

to Mary and succeeds in conveying a complete knowledge of all the facts about every member of [S]. Finally, and contrary to the view of at least some philosophers, we assume that what it is like for Smith to see red, Rp, is a property to which we are ontically committed; it cannot be construed as merely an ability or way of knowing an item which belongs to [S]. Hence, if Rp is physical, it must be included in [S] as a neurally constituted, although perhaps dispositionally characterised, physical trait. Having made all of these assumptions in favour of Jackson's argument, then, we can now consider how he might respond to the following situation.

We might imagine, for example, that in the distant scientific future an instrument has been devised which enables Mary to tap into Smith's experiences in such a way as to provide *completely reliable* information about what Rp is like for Smith [rather than merely what it is like for Mary to experience Smith's Rp *via the instrument*]. We shall name this remarkable piece of equipment the "Qualioscope". Neuroscience might be in a position to establish that the results obtained by using the qualioscope are indeed a completely reliable indication of Smith's Rp. Indeed, it is difficult to see how, from our present limited scientific standpoint, *all* the implications of neuroscience could be shown *not* to have such a capability.

The envisaged possibility serves to bring Jackson's position into sharper focus. What he must maintain is that if such results are possible, even in principle, then what Mary learns by using the qualioscope is a non-physical fact about Smith's experience. In order to justify this claim, however, he would have to resist the physicalistic alternative that Mary learns nothing; that the information conveyed via the qualioscope was already available. For according to reductive physicalism there are no facts about Smith's episode of seeing red other than the physico-dispositional facts already known to Mary. Since we have no reason to *presuppose* that there is anything non-physical involved in the qualioscope experiment, then, [i.e., we have as yet no further criteria for the physical] it seems that using the above line of reasoning commits Jackson to the a posteriori possibility that qualia should turn out to be physically conveyable, and hence physical items which Mary already knows by OPD.

If he *is* entitled to *presuppose* that all the physical facts are conveyable by OPD, however, his position will be distinct from Nagel's. Thus, for Nagel, we saw that facts are to be counted as physical just because they are objectively knowable. It might be argued, then, that for Nagel the qualioscope could only provide the information in question if physicalism were true. For if Mary has access to knowledge about Smith's Rp it follows that Smith's Rp is objectively knowable after all, and hence physical [any number of people might use the qualioscope as a standard piece of scientific equipment]. We would then have no logically prior reason even for assuming that such physical facts are not just some of the facts acknowledged by *current* physiology. Hence, in order to maintain a non-physicalist position with regard to qualia, Nagel would need to find some reason for supposing that the qualioscope *cannot work*, or cannot be *shown* to work even in the context of a complete grasp of current physiology. Jackson can resist this concession by citing some other criterion for the physical; he can insist that an item will only be counted as physical if it is conveyable by some specifiable means of demonstration [OPD] which does not include the use of a qualioscope. So even if the qualioscope works, and can be known to work, he will be entitled to insist that the information it conveys is non-physical. Even so, all the physical facts are, ex hypothesi, available to Mary by OPD. Hence, Jackson must still produce some explanation as to how he *knows* that there is anything left for Mary to discover, either via the qualioscope or by any other means. And with the information available so far, it is not at all clear what this discovery might amount to.

Conclusion.

What is apt to confuse the issue is that Jackson insists that Mary would have access to all the physical facts about other people while she is still in her room. This assumption invites speculation as to precisely what sort of evidence he is allowing Mary to have access to, and it is easy to be sidetracked into considering what sort of evidence should be sufficient for her to acquire knowledge of all the physical facts. The real problem with Jackson's argument rests with premise 2, however. What we actually need to establish is what, given all the physical facts about Smith's colour vision, Mary has yet to learn. If we already had a clear criterion for the

physical, which could provide justification for Jackson's experimental procedure, it would then follow that anything Mary has yet to learn must be non-physical. But in order to demonstrate that she would *have* anything more to learn he would have to establish at least that the physico-dispositional account of colour vision provided by a completed science would miss some of the facts. Unless he can find some other reason why this must be so, his argument simply begs the question it sets out to answer. His claim that even a completed science [or just a complete account of all the implications of current science] would fail to take certain facts about experience into account remains an a posteriori hypothesis.

We have seen that the further facts about qualia commonly invoked in an attempt to fill the gap in Jackson's reasoning fail to perform that function. Thus, it is claimed that qualia are distinct from physical items in that only the former have an *essential felt quality*, or are *essentially possessed by a conscious subject*. Neither of these suggestions affords any progress, however, since the essential attribute in either case might turn out on a posteriori investigation not to be essential after all. Thus, it might turn out that the felt quality we refer to as a quale is just a physico-dispositional trait belonging to [S] after all. Similarly, it might turn out that what we thought to be essentially possessed by a conscious subject is just a physico-dispositional trait [belonging to [S]] which can also be possessed by an unconscious subject. What we need at this stage, then, is some further fact about qualia which ensures that this a posteriori possibility cannot turn out to be a fact. Or, in more general terms, we need to be able to cite some characteristic of qualia which all physical properties are already known to lack.

In pursuit of this characteristic we turn next to Kripke's modal argument. If it is successful it will show that the characteristic possessed by qualia, but not by any physical properties, is that qualia are not *necessarily* identical with any members of [S]. All occurrent *physical* properties, in contrast, would have to be necessarily identical with members of [S]. If the argument is successful, then, Kripke will be entitled to infer that qualia are not even members of [S] as a matter of fact.

Chapter VI

KRIPKE'S INTUITION.

In the discussion so far, we have reduced the physicalist's position with regard to qualia to the starkly undifferentiated claim that:

[QP] There are no occurrent irreducibly non-PPD qualia.

Hence, for the physicalist, qualia can only be said to be occurrent items if qualia-discourse is construed as discourse about physico-dispositional items which are acknowledged, at least in principle, by a future, completed physical science. In accordance with our construal of reductive physicalism, however, we wish to be more specific. Thus, the thesis being evaluated is that there are no occurrent qualia in addition to the properties already accepted by QP and QD alike as belonging to the set [S] of items to which *current* physical theory is ontically committed.

Notwithstanding the inverted spectrum possibility in the case of the *exhibition* of relatively simple dispositional traits, it was shown that we have, at present, no prima facie grounds for dismissing the proposed characterisation of qualia as physiologically [or *neurally*] grounded traits which can be characterised in either paradigmatically neural terms or dispositional terms.

In accordance with the first option, the physicalist's thesis would have to be that an introspectible property Rp, of the epistemically topic-specific, but neurally topic-neutral, type [p], is just some neural property N of type [N]. Multiple instantiation by types [N] .... [N_N] is ruled out, since if the identity thesis we are considering is correct, all introspectible types are also identical with specific neural types. Since it seems clear that the introspectible property Rp *is* of a specific type [Rp], our physicalism ordains that according to this option type [Rp] must itself be some specific neural type [RP]. And we argued in chapter IV that it would be a matter for a posteriori investigation to discover whether properties of type [Rp] even *co-occur invariantly* with neural properties of any single type [RP].

The second option entails that Rp is of no particular neural type [N], except insofar as its is of some type [N] ... [N_N] which has certain dispositionally characterised properties. We argued that any occurrence of Rp would at least have to be characterisable as the occurrence of some neural state or other which would, in standard conditions, lead to the exhibition of Rp-appropriate dispositions. Here, the exhibition of a disposition would amount to the occurrence of behavioural traits symptomatic of that disposition. And this entails that there is one or more neural state *type* which satisfies this dispositional requirement invariantly. Thus, introspectible property Rp will be topic-neutral with regard to neural types, except insofar as each neural state will share the property of disposing the subject to exhibit Rp-appropriate behaviour in standard conditions. The onus on QD is then to establish that there are no neural states which satisfy this condition invariantly in standard conditions. Again, this is a matter for further a posteriori investigation and therefore the case has not been resolved either way. Even if it had been resolved in favour of the reductive physicalist, however, a relation of *supervenience*, but not *identity*, would have been established. Further investigation or argument is required to establish that invariantly co-occurrent types are in fact identical.

Further, the knowledge argument leaves the physicalist's proposal unchallenged. For although the intuition seems compelling that qualia are essentially felt qualities, or essentially possessed by a conscious subject, while physical items are not, that intuition might nevertheless owe its air of compulsion merely to the limited scientific perspective at present available to us. From the perspective even of a more comprehensive command over the implications of *current* science, it might turn out that the intuition is misguided. Thus, it might turn out that the felt qualities which are possessed by a conscious subject need not be. To assume that they have this characteristic essentially is to beg the question as to whether they are physical items in favour of [QD]. At the very least, we can say that a rebuttal of physicalism by recourse to the knowledge argument requires that we *presuppose* that introspectible properties are not just members of [S]. In general, we can say that no characteristic of qualia has been cited which precludes the possibility of physical reduction in principle. Whether qualia are physico-dispositional items remains to be

established by further a posteriori investigation, by gaining further information about the nature and properties of physical items.

Kripke acknowledges the possibility of an a posteriori physical reduction of introspectible properties as his starting point, but argues that if even a *token*-identity relation between qualia and physical properties or states obtains, it must be a relation of *necessary* identity between the two. One way in which he expresses this conviction is by saying that any identity *statement* in which the referents of the referring expressions are *rigidly designated* must be a necessary truth; epistemically, a token pain is picked out topic-specifically by direct introspection without appeal to intermediary properties. Hence, if an identity statement such as "this token pain is identical with a token C-fibre stimulation" is true, it is necessarily true. In the case of phenomenal properties, such as pain, however, we have the strong intuition that such statements are *not* necessarily true. From the observation that they do not *appear*, intuitively, to be necessarily true, then, he is able to infer that they are probably not even true. At least, some explanation will be required as to how they might be true in the face of his intuition, and he regards the prospect of producing such an explanation to be a considerable challenge. Taking token identity to be the thesis in question, his argument can be summarised more fully in the following way [As expounded in Kripke, 1980]. If it succeeds, it will follow that no *type*-identity thesis can be true either.

1. A *rigid designator* is a designator which designates the same object in all possible worlds.

2. A *necessary truth* is a truth which obtains in all possible worlds.

Therefore, from 1 and 2,

3. An identity statement which involves rigid designators is, if true, true in all possible worlds [i.e., necessarily true].

4. "This pain" and "this C-fibre stimulation" are both rigid designators.

Therefore, from 3 and 4,

5. If "this pain is identical with this C-fibre stimulation" [I] is true, it is necessarily true.

6. But it seems intuitively that I is not necessarily true.

Therefore, from 5 and 6,

7. It seems intuitively that I is false.

Therefore, from 7, and unless some other adequate explanation for the intuition in 6 can be found,

8. This pain is not identical with this C-fibre stimulation.

This argument depends on premises 1, 2, 4 and 6, being true, and inferences 3, 5, 7 and 8 being valid. We might therefore proceed by trying to clarify each of these points in turn. Since we are not particularly interested for present purposes in the semantic aspects of Kripke's argument, however, it will be helpful to reformulate the argument without reference to rigid designators, and necessary *truths*. Reformulated appropriately, denoting this pain as P, and this C-fibre stimulation as CFS, the argument runs as follows.

1. If P is identical with CFS it is necessarily identical.

2. It seems intuitively that the modal proposition [M], that P is not necessarily identical with CFS, is true.

Therefore

,

3. It seems intuitively that P is not identical with CFS.

Therefore, in the absence of any other satisfactory explanation for the intuition in 2,

4. P is not identical with CFS.

Clearly, this argument depends crucially on premise 2; that we have a strong intuition to the effect that P is not necessarily identical

with CFS. Our first task, then, will be to attempt to find out exactly what this intuition is supposed to amount to. The first important point here is that if Kripke's argument is to be construed as a *modal* argument we must at least be able to make sense of his concept of *necessity*. Secondly, if the modal intuition is to carry any weight, we shall need to understand what it amounts to have such an intuition. In the ensuing discussion, therefore, we shall draw a distinction between the *modal proposition* [M] on which the argument is founded, and the *intuition* [I] that [M] is true.

Kripke's Concept of Necessity.

In an attempt to clarify the concept of necessity invoked by the modal proposition we shall look firstly at a brief passage in Kripke's account which is intended to provide a summary of the argument based on that proposition. Since, at present, we are interested only in interpreting the notion of necessity itself, and the intuition is assumed for the purpose of the argument to be present, subsequent features of the argument will simply be accepted uncritically. Although Kripke evidently finds his references to God and his powers helpful, we shall attempt to capture the spirit of each premise without them. Also, we shall assume that since what we are interested in finding out is whether or not *our own occurrent pain P* is a token physical state or property at all, we shall restrict the argument to such a pain. We are not interested primarily in the possibility that Martians might have pains without having any C-fibres, for example, or even that other occurrent pains might not be other episodes of CFS, except insofar as such possibilities carry implications for our own occurrent pain P[36].

---

36. This assumption allows Kripke's argument to operate on the basis of a weaker modal intuition. Thus, while an occurrent pain P might be identical with a C-fibre stimulation, and necessarily so, it does not follow that all other pains, even in all other possible worlds, will also satisfy these conditions. We need not insist at this stage in the argument against *token*-identity that other occurrent pains, or non-occurrent pains which God might have created, are identical

Accordingly, we shall attempt to make sense of the concept of necessity at least initially without reference to any pains other than P, Martian or otherwise. The term P will be employed to refer to my occurrent pain at time t. The argument we are about to examine is intended to establish that the modal intuition is present, but for now we wish to understand the modal proposition itself. The following question offers an initial insight into the way Kripke intends to expound his concept of necessity, and [M] itself.

> What about the case of the stimulation of C-fibres? To create this phenomenon, it would seem that God need only create beings with C-fibres capable of the appropriate type of physical stimulation [and also, presumably, bring it about that they be so stimulated]. Whether the beings are conscious or not is irrelevant here. [1980, p 153] (My parentheses)

It seems fairly clear that the following redraft captures the spirit of this remark, at least approximately. We shall assume that the CFS referred to is the particular token physical state or episode purported to be identical with *this particular pain* P. Further, since the modal relation in question is the subject of Kripke's intuition in premise 2, we shall consider the following as the intended propositional *content* [M] of that intuition. Thus, the intuition expressed in premise 2 might be that:

[M]1 If the CFS occurs then the CFS occurs, irrespective of whether it is P.

On this interpretation, the first part of [M]1 is trivially true, so that on its own it could legitimately be disregarded as a premise; logical truths cannot play any useful part as premises in a logical argument, since they are implicitly assumed anyway. So this leaves the second part to be considered. The premise is that this logical

---

with C-fibre stimulations, even if the particular token pain P he created is [see Carruthers, pp 152 - 3 for this point]. It might turn out from Kripke's further considerations that the latter claim does follow from the former, but we do not need to presuppose here that it does.

truth obtains whether or not some other proposition p [i.e,., "this occurrent pain is the CFS"] is true. On the assumption that the premise is not intended as the mere affirmation of a logical truth, then, we might assume that it is intended to affirm that p is not entailed by that logical truth. But nothing is entailed by a logical truth other than another logical truth, so that cannot be intended either, since p is not a logical truth [if it were, there would be no possibility of establishing that p is false]. The only remaining interpretation of the premise is therefore that it is possible for the CFS to occur even if p is false. Thus:

[M]2 The CFS can occur even if it is not P.

But, in the absence of independent support, this *presupposes* that the CFS is not P, and we are not entitled to make that presupposition without begging the very question at issue. The conclusion that the CFS is not [likely to be] P is to be arrived through Kripke's *modal* proposition, which has yet to be brought into play. Consequently, the strongest permissible interpretation of the propositional object of the intuition in 2 at this stage in the argument must be that:

[M]3 *It is possible* that the CFS can occur even if it is not P.

And, unless this is taken to be an expression of the modal relation itself, this amounts just to saying that at this stage we do not *know* whether the CFS and P are identical. In order to avoid reducing it to the mundane observation that the identity relation between the CFS and P can, if it holds, has yet to be established, then, we will have to read it as the declaration of a modal relation. Thus:

[M]4 It is "*modally*" possible that the CFS can occur even if it is not P.

And whatever this "modal" possibility might amount to, we must assume that for Kripke's purposes it entails that the CFS and this pain are not *necessarily* identical. At this stage, then, we must see whether proposition [M]4 can be further clarified as the subject of an intuition.

> God had to do some work, in addition to making the man himself, to make a certain man the inventor of bifocals; the man could well exist without inventing any such thing. The same cannot be said for pain; if the phenomenon exists at all, no further work should be required to make it into pain. [p 154]

Here, the purported comparison with the inventor of bifocals seems irrelevant to what Kripke has to say about pain. Thus, if God created a certain man, no further work would be needed to make him that man. But by the same token, if he created the inventor of bifocals [a man who had the property of being the inventor of bifocals], no further work would be needed to make him the inventor of bifocals. This is exactly parallel with the point summarised in premise [M]1; in this case, in order to create "the inventor of bifocals", God need do nothing other than create a man with the required capability and bring it about that he deploys it. As far as *this pain* is concerned, therefore:

[M]1' P is nothing more than P.

or:

If P occurs then P occurs.

Under either interpretation we can see that the premise need not detain us, since it is just another logical truth. If on the other hand, Kripke is trying to say something about the CFS here, we might assume that he is saying that:

[M]2' P can occur even if it is not the CFS.

And while this is logically distinct from [M]2, it is no more enlightening with regard to the nature of Kripke's *modal* statement. So as with premise [M]2, we must put premise [M]2' on hold until we have some modal proposition with which to support it. So far, we have made no progress beyond a modal interpretation of [M]4, and as yet we have no indication of what the modality featuring in that proposition amounts to.

But now Kripke produces a further account of the crucial modal relation.

> It would seem, though, that to make the C-fiber stimulation correspond to pain, or be felt as pain, God must do something in addition to the mere creation of the C-fiber stimulation; He must let the creatures feel the the CFS as *pain*, and not as a tickle, or as warmth, or as nothing. ....if so, the stimulation could exist without the pain. [pp 153 - 4]

An initial stab at extracting the propositional content of this intuition might be that, intuitively:

> [M]2" *Feeling* the CFS as this pain is something more than just the CFS occurring; the CFS could occur without being felt as this pain.

But this will clearly not do either, since the token-identity thesis under consideration is not that the CFS is *felt* as this pain, but that it just *is* this pain. In order to make the premise relevant in this context then, it might be redrafted as:

> [M2]" This pain occurring is something more than just the CFS occurring; the CFS could occur without the pain.

And this is clearly no better than the original [M]2; all it relevantly says is that the CFS could occur without P occurring, and in order to avoid *presupposing* that the CFS is not P, it must be interpreted modally as in [M]4. Thus:

> [M]4' It is "*modally*" possible that the CFS can occur even if P does not occur; i.e., they are not *necessarily* co-occurrent.

The challenge is then to understand what this concept of necessity amounts to without having to resort to premise [M]2". Thus, if it is to be interpreted modally, it will have to be supported independently of the intuition in [M]2" that P is something more, or other than, the CFS; otherwise it simply begs the question of whether the latter is true, and the modal premise becomes redundant. So we might initially construe Kripke's concept of necessity as the *impossibility of a state of affairs failing to occur*. His modal claim is then just that the pain *could*, in his modal sense, occur without the CFS also occurring, and vice versa.

We can now turn to other references in Kripke's account of the modal proposition he has in mind in an attempt to clarify it further. Perhaps a suitable starting point would be with his references to the Cartesian intuition. He says:

> Descartes, and others following him, argued that a person or mind is distinct from his body, since *the mind could exist without the body*. He might equally well have argued the same conclusion from the premise that the body could have existed without the mind. [pp 144 - 5]

Reformulated in terms of this pain P at time t, and a particular episode of CFS, we can see that this Cartesian intuition is just another version of [M]2". But Kripke introduces a modal element into the account when he indicates the intended reading in the following passage.

> Let 'A' name a particular pain sensation, and let 'B' name the corresponding brain state, or the brain state some identity theorist wishes to identify with A. *Prima facie*, it would seem that it is *at least logically possible* that B should have existed [Jones's brain could have been in exactly that state at the time in question] without Jones feeling any pain at all, and thus without the presence of A. [p 146]

Thus, the intuition is that if the identity relation obtains at all [which must remain a possibility at this stage in the argument], it is a contingent fact, rather than a logical necessity. He refers again to the Cartesian intuition as the intuition that:

> A can exist without B, that B can exist without A, that the correlative presence of anything with mental properties is merely contingent to B, and that the correlative presence of any specific physical properties is merely contingent to A. [p 148]

On this reading, then, we can say that the intuition in question is not that this pain and the CFS are distinct, but that logically, *either can, or could, occur without the other*. This is what gives the intuition its modal import. It seems that the two phenomena are not *necessarily* co-occurrent. The inference that the two seem to be distinct phenomena is to be drawn on the basis of the intuition expressed in [M]4' that they *need not co-occur*, in conjunction with

his further thesis concerning the necessity of identity, as expressed in premise 1.

Before going on to consider what *having the intuition* to this effect might amount to, it will be helpful to refine the account of the modal proposition in question with one further rider. Thus, whereas in the case of heat/molecular motion we already know the identity relation to obtain, we do *not* already know that this pain P is the CFS. In the scientific example, then, Kripke's modal proposition [M]4 is that, logically, the identity relation *might not have obtained, even though it does*. The crucial point is that although [M]4 is false, because since the identity relation obtains it does so necessarily, he still has the *intuition* that logically it might not have obtained at all. In this case, then, the intuition is false, and will need to be "explained away". In the P/CFS case, however, we do *not* already know that the identity relation does not obtain. But it would be disappointing if the modal proposition in this case turned out to be just that we do not yet know whether P is CFS. Hence, if the parallel with the scientific modal proposition is to be maintained, he must be saying that even if it were found to obtain he would still have the intuition that, logically, it *might not have done so*. In both cases, it *seems* that logically the identity relation might not have obtained, even if it does necessarily obtain. In accordance with Kripke's exposition of the modal proposition, then, we can characterise it as:

> [M] Irrespective of whether the identity relation does obtain, and necessarily so, *it might not have done so.*

Or, in order to bring the fact that it is false into sharper focus, we might reword it as:

> [M] Irrespective of whether the identity relation obtains, and necessarily so, *it does not necessarily obtain.*

Since [M] in this form is patently false, then, it is imperative that he find some way of explaining away his intuition that it is true. The question then is how he proposes to explain away the intuition that [M] is true in the scientific case without having to sacrifice the necessary identity relation in the process, and why

the same sort of explanation cannot be applied also to the P/CFS case.

If the modal proposition is patently false, then, what are we to make of Kripke's modal *intuition*? His own explanation of the intuition in the scientific case is that in fact it amounts just to the observation that it is *epistemically possible* that [M] is true. We can now see what this explanation amounts to.

Kripke's Scientific Essentialism.

Kripke maintains that there is a crucial distinction between necessary identity relations in science, which can be known to obtain only a posteriori, and purported identity relations between a mental phenomenon and a physical phenomenon. The distinction is based on the epistemic observation that:

> Pain.... is not picked out by one of its accidental properties [as, for example, heat might be picked out by the *sensation* of heat]; rather it is picked out by the property of being pain *itself*, by its immediate phenomenological quality. Thus pain, unlike heat, is not only rigidly designated by "pain" but the reference of the designator is determined by an essential property of the referent. Thus it is not possible to say that although pain is necessarily identical with a certain physical state, a certain phenomenon can be picked out in the same way we pick out pain without being correlated with that physical state. If any phenomenon is picked out in exactly the same way that we pick out pain, then that phenomenon *is* pain."
> [pp 152 - 3] (My parentheses)

The essential point here with regard to the scientific case appears to be that there are, indeed, identity relations which obtain [necessarily] but can only be *known* to do so a posteriori. Although it is true, and hence for Kripke necessarily true, that heat and molecular motion are one and the same physical phenomenon, there is a sense in which the identity relation can seem to be only contingent. Thus, since heat is typically picked out epistemically via certain of its properties [it produces a *sensation* of heat in human beings, for example], it seems possible that the *cause* of those properties might have been something other than heat

[molecular motion]. This possibility obtains in virtue of the *topic-neutrality* of the properties picked out, with respect to their physical cause; whatever is causing this sensation of heat, for example, might not have been molecular motion. For we can imagine the physical nature of the world being other than it in fact is, and that in such a situation different causal laws obtain. Molecular motion might, in accordance with such laws, have caused what we now take to be a different sort of sensation; a dull ache, a tickle, or even no sensation at all. By the same token, the cause of the sensation of heat might have been something other than heat/ molecular motion. Thus, in such cases, our intuition that the necessary identities of science are only contingent might be "explained away" as our ability to see that the *sensory effects* of the physical phenomena involved might have been other than they in fact are. Hence, the epistemic explanation of the illusion of contingency is at least plausible in such cases.

What Kripke is saying in the scientific case is then that although:

> [A] Heat is necessarily molecular motion,

the epistemic facts are such that:

> [B] The phenomenon [heat] which causes the sensation of heat [and through which we pick out heat epistemically] might have caused some other type of sensation instead, if the physical laws had been different.

And in view of [B], we can say that there is a certain illusion that [A] is false, and that having his modal intuition amounts just to succumbing to that illusion. The illusion occurs because it is possible to think of heat *topic-neutrally*, as whatever causes the *sensation* of heat. By failing to recognise that we are thinking of heat topic-neutrally in this way, we might suppose erroneously that heat itself, topic-*specifically*, is not necessarily molecular motion. Once we see that we have been thinking of heat topic-neutrally, however, it is possible to explain away the illusion without having to sacrifice the modal proposition [A]. Heat is necessarily molecular motion, but the sensation of heat through which we identify heat epistemically might not have been caused by

heat. So the patent falsehood expressed in [M] can be translated plausibly into the fact that:

> [M] Irrespective of whether the identity relation obtains, and necessarily so, it is not necessarily the case that the *sensation* of heat is caused by heat/molecular motion.

This fact is then contrasted with the case of mental phenomena. When the phenomena which are supposed to be identical with physical phenomena are themselves the introspectible *sensory effects* of a particular phenomenon, those effects can be picked out epistemically and topic-specifically without the intervention of any properties whatever. If something feels like a pain it just is a pain, *since the phenomenon of pain is nothing other than the sensory experience itself.* Hence, according to Kripke, the epistemic explanation of how an identity between a pain and a CFS might seem to be only contingent is simply unavailable. If a pain *were* identical with a CFS, the appearance of contingency could not amount to the possibility that the pain might not have caused the sensations through which we pick out pain epistemically; for the sensation just is the pain. There are therefore no intervening properties in terms of which to formulate a topic-neutral explanation of the appearance of contingency. In this case, then, the correlate of [M]:

> [M]' Irrespective of whether the identity relation obtains, and necessarily so, it is not necessarily the case that the *sensation* of pain is caused by P/CFS

is simply not available, since the pain P just is the sensation.

Kripke's position, then, is that whereas the illusion of contingency can be explained away in cases involving a necessary identity between tokens of heat and molecular motion, the same explanation cannot be applied to cases involving tokens of pain and C-fibre stimulation. He confirms our interpretation by explaining that:

> In the case of molecular motion and heat there is something, namely the sensation of heat, which is an intermediary between the external phenomenon and the observer. In the mental-physical case no such intermediary is possible, since here the physical phenomenon is supposed to be identical with the internal phenomenon itself. Someone can be in the

> same epistemic situation as he would be if there were heat, even in the absence of heat, simply by feeling the sensation of heat; and even in the presence of heat, he he can have the same evidence as he would have in the absence of heat simply by lacking the sensation S. [pp 151- 2]

And from these considerations he infers that since we do have the same intuition in both the scientific and the pain/CFS case, the intuition must be true in the latter case; there is apparently no way of explaining away our intuition that P might not have been CFS, and so, in the absence of any other sort of explanation for the intuition, P really might not have been CFS. But if P is not *necessarily* CFS it cannot even *be* CFS [since if the identity relation obtained it would obtain necessarily].

One limitation of the Kripkean intuition here is worth underlining. Thus, Foster's Cartesian appeal takes the premise [M]4' further. He argues that even if a neural event N is identical with a pain event P,

> ...we can surely envisage a counterfactual situation in which exactly the same neural event occurs in Smith's brain at t [its identity as N being fixed by its physical properties, its brain location, and its causal origins], but in which, with a suitable change in psychophysical laws, Smith does not have a pain experience at t. But if we retain both these intuitions, we are forced to conclude that P and N are numerically distinct. [Foster, 1991, p 135].

But the modal proposition was that:

> [M]4' It is "*modally*" possible that the CFS can occur even if P does not occur; i.e., they are not *necessarily* co-occurrent.

Now even if we construe Foster's statement somewhat charitably as a restatement of the *modal* intuition, he is still seriously in error. Thus, his reference to psychophysical laws is simply inappropriate. For if, ex hypothesi, N and P are *identical*, there are no psychophysical laws to invoke. If there were, the intuition concerning P/CFS, or P/N, could be explained away in exactly the same way as we explained away the intuition in the heat/molecular motion case. So the whole point of Kripke's challenge is that we

cannot allow that a suitable change in such laws would render the two not identical, since if they are identical the identity is necessary, and there are therefore no such laws to be considered. What we can allow is just that the *observable effects* E produced by N [an increase in bodily adrenalin, for example], might have been produced by some event other than P if the psychophysical laws had been different. Thus, we might allow that *the particular cause of E at t* could have been something other than N. But if this were offered in support of Kripke's intuition, it would be self-defeating. For in that case it would also *explain away* the intuition without the need to sacrifice the identity relation itself; precisely what Kripke wishes to avoid. Kripke's argument is that the intuition is present, but there is no plausible way of explaining it away.

There is also a second limitation worthy of note. Kripke explicitly endorses the epistemic explanation of his "illusion of contingency" in cases involving the supposed identity relation between heat and molecular motion, for example. It seems clear, then, that any objection to the above explanation of the illusion of contingency in the case of pain and the CFS must insist that the epistemic situation in which he has his intuition is not of that sort. But Geoffrey Madell, for example, expresses surprise that anyone should find any epistemic explanation even *intelligible* in the case of pain [p 95]. Thus, while we can readily concede that identity relations which obtain in science or mathematics might seem to be only contingent, as a consequence of our present limited physical knowledge or understanding [or in virtue of the epistemic topic-neutrality of our concepts of scientific phenomena], the same cannot be said for this pain sensation P and CFS. But his reason for making this distinction is that in the latter case:

> Far from one seeing any cause to doubt it [i.e., seeing that it is just *epistemically* possible that it should turn out not to obtain, or that it might not have obtained], it becomes ever more clear that the suggested identity between [this pain sensation] and [the CFS] is *incomprehensible* [p 95]. (My parentheses)

But Kripke's modal intuition cannot be just that the suggested identity relation is incomprehensible *per se*. If he were claiming this, his attempt to show that there is no plausible way of

explaining the intuition of contingency away would be entirely redundant. For instead of claiming [modally] that it seems that the relation *might not obtain*, he would be in a position to say rather that the identity thesis is false just because it is unintelligible; and this is not the modal consideration he has in mind. What he must presumably insist on instead is that the *epistemic explanation* is incomprehensible, or perhaps just patently false. Within the context of his argument, the identity thesis which Kripke claims to be false cannot itself be incomprehensible. If it were, there would be no point in trying to show that it is not *necessarily* true.

Finally, we might note that if the epistemic comparison offered by Kripke is correct, there is a sense in which the intuition in the scientific case must be different from that in the pain case. For if the epistemic situation in which the scientific intuition occurs includes the detection of heat via one of its sensory *effects*, while in the case of pain it does not, the two intuitions are, strictly, occurring in different epistemic situations. This is a minor point, however, and is of no real concern. For the point is just that in both cases there is a strong inclination to believe that the identity relation is not necessary, and in that respect they are the same.

Possible Objections to the Epistemic Disanalogy.

In order to evaluate *any* proposed objections to Kripke's argument we shall need to set certain ground rules which have the following function. We can understand at least in broad terms both his modal proposition and his intuition that it is probably true. In the case of scientific identities, between heat and molecular motion for example, we can understand the modal proposition as the negation of [A] (p 204); it is the proposition that even if the identity obtains[37]:

---

37. We should recall that although this rider renders the overall proposition false for Kripke, the *intuition* is that it is true.

-[A] Heat is not *necessarily* molecular motion.

We can also understand what it amounts to for Kripke to have the *intuition* that -[A] is probably true; it is simply a strong inclination to believe that it is true. Similarly, then, we can understand both the modal proposition and the intuition relating to pain and C-fibre stimulation. In this case, the modal proposition is as indicated on page 202, that even if the identity obtains:

[M] [This token] pain is not *necessarily* a C-fibre stimulation.

The intuition that [M] is probably true is then just the strong inclination to believe that it is true.

In the first case, Kripke provides a way of explaining the intuition away without having to sacrifice the necessary identity relation itself. Thus, he explains [see p 204] that we have the intuition because:

> [B] The phenomenon [heat] which causes the sensation of heat [and through which we pick out heat epistemically] might have caused some other type of sensation instead, if the physical laws had been different.

Kripke's problem is then that no such explanation is available in the second case, from which he infers that probably no satisfactory explanation whatever is available.

Our problem, then, is to determine at this stage what *would* constitute a satisfactory explanation. For if we agree that pain just is the sensation, there are indeed no intermediary properties through which pain is picked out epistemically, so that the strict corollary of [B] really is unavailable. In the absence of any further guidelines from Kripke, then, it seems that an explanation will be satisfactory just if it is a plausible explanation of how we come to have the strong inclination to believe that, even if this pain is identical with a CFS, [M] is true. For in the first case we have no indication that *Kripke* requires any more than this. We might just add that such an explanation would have to refer just to the *epistemic* facts, as does Kripke's explanation for heat and molecular motion.

From our considerations in chapter III, however, it seems far from clear that *all* epistemic explanations for the modal intuition must be patently false. Nagel, for example, suggests the following explanation:

> A theory that explained how the mind-brain relation was necessary would still leave us with Kripke's problem of explaining why it nevertheless appears to be contingent. That difficulty seems to me surmountable, in the following way. ..... To imagine something perceptually, we put ourselves in a conscious state resembling the state we would be in if we perceived it. To imagine something sympathetically, we put ourselves in a conscious state resembling the thing itself. .... When we try to imagine a mental state occurring without its associated brain state, we first sympathetically imagine the occurrence of the mental state.... . At the same time, we attempt to perceptually imagine the non- occurrence of the associated physical state, by putting ourselves into another state unconnected with the first: one resembling that which we would be in if we perceived the non-occurrence of the physical state. [Rosenthal, p 428, footnote 11]

In terms of our considerations in chapter III, we can express the same point in the following way. Suppose that P, this sensation of pain, is in fact identical with the CFS and, as Kripke would then insist, necessarily so. Nevertheless, we can explain the intuition that the identity relation seems to be only contingent once we recall that there are two quite different types of epistemic situation we can be in and yet discern an occurrence of the CFS/sensation of pain. Firstly, there is the first-person perspective; the epistemic situation Eq in which we are able to discern [topic-neutrally with respect to the type or token of physical phenomenon being thus discerned] in introspection that we have an occurrent pain sensation P, *even though we might not know that P is in fact the CFS.* Secondly, there is the third-person perspective; the epistemic situation EP which we would be in when observing the brain from the outside, in paradigmatically scientific fashion. In the latter epistemic situation, we might be able to discern topic-*specifically* that we are undergoing an episode of CFS, *even though we might have insufficient information to determine that this CFS is identical with the pain sensation.* Hence, the illusion of contingency can be explained as the *epistemic* possibility that the pain sensation, as discerned in Eq, might not have been, or

might not turn out to be, identical with the CFS, as discerned in EP. It becomes intuitively plausible to suppose that the pain sensation and the CFS might not even be invariably co-occurrent, as Descartes pointed out. We might even find it plausible to suppose, as McGinn has,[38] that we are cognitively incapable of ever understanding how we can have two such diverse perspectives on one and the same physical phenomenon. In any case, it seems unsurprising that this pain sensation, as picked out topic-neutrally in Eq, might turn out not to be invariably co-occurrent with the CFS, and therefore find it natural to suppose that the identity relation is only contingent, even if it is in fact necessary.

It is important to recognise that the above epistemic explanation does not depend on CFS being detected only *topic-neutrally* in the situation EP. For, as we saw in chapter V, the reductive physicalist will insist that all of the occurrent properties and states will be fully intelligible and epistemically available in the third-person perspective, at least in principle. What our explanation requires is just that there is another epistemic situation Eq in which CFS is discerned only topic-neutrally. If it could be established that the pain discerned in introspection seems to have some characteristic X which CFS does not, the position would be different; for then we would have to explain why P, but not CFS, seems to have X. Thus, Kripke might insist that having a pain [sensation] is *more* than just discerning any physical phenomenon topic-neutrally in introspection, and therefore infer that our explanation of the intuition is inadequate. But so far no such characteristic X has been convincingly cited. Hence, until such a characteristic can be found, the epistemic explanation seems sufficient. It seems to Kripke that the identity relation P/CFS is not necessary, but this is just because P is discerned only topic-neutrally in Eq, and therefore what is discerned in Eq might not be, or might not have been, a CFS.

We might reasonably suppose that this explanation conforms to Kripke's own expectations. For although we are saying that there are no intermediary introspectible properties through which CFS is

---

38. Colin McGinn, 1994.

discerned in Eq, it is epistemically possible to discern CFS in introspection [as this pain] without knowing that it *is* CFS. Hence, there is a clear parallel between the sort of explanation offered by Kripke in the first case, and our own explanation in the second.

In view of this explanation, we can now suggest that Kripke's appeal to an analogy between the case of heat and molecular motion, and that of this pain and the CFS, is misleading. For while it might indeed be the case that a similar illusion of contingency obtains in both cases, there is no need to appeal to *exactly* the same sort of *explanation* for both. The explanation for heat and molecular motion is that heat is picked out epistemically through the recognition of certain of its *properties*. In contrast, the explanation for pain and CFS is just that this pain is discerned in an epistemic situation quite unlike that in which CFS, *per se*, is typically discerned. It is the availability of these distinct types of epistemic situation which explains how the latter identity relation can seem to be only contingent. Kripke misleads us into expecting the explanation to be of *exactly* the same type as that used to explain the illusion of contingency in the case of heat and molecular motion. Thus, he seems to assume that since there are no *intermediary properties* in the case of pain, no satisfactory epistemic explanation can be provided.

A Metaphysical Version of Kripke's Intuition.

One way in which Kripke's position might be defended against our epistemic explanation is by stating an apparently stronger version of the modal proposition in question. Thus, instead of claiming just that the same sort of illusion is present in both cases, he might argue that in the case of heat we have a further intuition which is simply *absent* in the case of pain. We should note that in this case the *type*-identity thesis is in question. He suggests that:

> It certainly represents a discovery that water is H_2O. We identified water originally by its characteristic feel, appearance and perhaps taste. If there were a substance, even actually, which had a completely different atomic structure from that of water, but resembled water in these respects, would we say that some water wasn't H_2O? I think not. We would say instead that just as there is a fool's gold there could be a fool's water; a substance which,

> though having the properties by which we originally identified water, would not in fact be water. [p 128]

The intuition here, then, is that if something were to have a chemical composition different from that of water it would not *be* water. Similarly, if a physical phenomenon which produced a sensation of heat were not molecular motion, it would not be heat. Hence, the new intuition appears to be stronger than the original Cartesian one. It amounts to the more direct modal claim that we have an intuition that heat *is*, in a metaphysical sense, necessarily, or essentially, molecular motion, even though there are epistemic situations in which it might seem not to be.

The claim would then be that such an intuition is simply *missing* in the case of sensations of Pain. Thus, suppose that we were to discover that we had been wrong all the time about the identity of Pain and the CFS, construed as *types* of phenomenon. It turns out on investigation that Pain is either a physical phenomenon of some other type or no physical phenomenon at all. The suggestion would then be that the phenomenon we have picked out as *this* Pain would still be counted as a Pain; the intuition that we would take it to be something other than Pain is simply missing.[39] Hence, it appears that we now have a stronger reason for inferring that the type-identity of Pain with CFS is not necessary.

In the light of our epistemic explanation for Kripke's original [Cartesian] intuition, however, we can see that the above suggestion lacks the required import. If Pain is identical with CFS, but we employ the term "Pain" to refer only *topic-neutrally* to whatever we discern in introspection in Eq, the identity will nevertheless be metaphysically necessary. But the fact that we do not have the intuition that if Pains, as discerned in introspection, turned out not to be CFSs, they would not be Pains, can be explained on purely epistemic grounds. We can say that the illusion of contingency in this case amounts just to our ability to envisage Pains, as

---

39. For this suggestion, see, for example, George Bealer, p 368 - 74

discerned topic-neutrally in introspection, turning out not to be CFSs.

A similar possibility can be envisaged in cases of scientific identity. Thus, suppose that we know of a type of physical item which exhibits properties X, Y and Z, but that we *do not yet know* its molecular composition. Suppose, in fact, with George Bealer, that we are able to pick out samples of CFS just by probing the body with the scientific instruments available at the time and finding that they all exhibit properties X, Y and Z. We might *suspect* that these samples of CFS all have at least 74,985,262 functionally related non-conscious parts [p 371]. What, then, would we make of a sample of CFS which has been identified in the usual way, but which turns out on further investigation to have *fewer* parts? In such a situation it seems clear that the proper conclusion would be that samples of CFS do not all have at least 74,985,262 parts. The sample with fewer parts would still be a sample of CFS. Hence, as in the case of Pain, the intuition that if this sample of CFS had fewer parts we would infer that it is not a CFS after all is simply missing. And the reason for this is that what we refer to as "samples of CFS" are just, topic-neutrally, any samples exhibiting the relevant scientific properties. Irrespective of how much we know, even, we can still say that if the term "CFS" is employed *topic-neutrally*, to refer to whatever physical phenomenon has properties X Y and Z, our intuition is that if a sample picked out as such turned out not to have a certain number of parts we would still count it as a CFS.

Of course, we *could* employ the term 'CFS' differently, to *mean*, at least in part, 'physical items having at least 74,985,262 parts'. If 'CFS' is used in this way, it does seem intuitively plausible that any sample found to have fewer parts would not be a sample of CFS. But the same applies to Pain. If we were to employ the term "Pain" to refer specifically to CFS, the discovery that this particular item discerned in introspection, and hitherto referred to as a sample of Pain, is not a sample of CFS would lead us to infer that this sample is not a Pain after all. So in that case the intuition cited by Bealer is not missing at all. The relevant point is just the semantic one that 'Pain' can be used in two distinct ways. Hence, the intuition that if 'this Pain' turned out not to be a CFS we would still call it a pain fails to indicate that this Pain is

not necessarily a CFS. The intuition which Bealer claims would be missing in the Pain/CFS case would also be missing in the CFS/n-parts case, just if we had resolved in advance to use the term CFS topic-neutrally with respect to the number of its constituent parts.[40]. Again, the objection to this account would have to be that discerning Pain in introspection amounts to more than just discerning CFS topic-neutrally in Eq; but at this stage such an objection has yet to be substantiated.

We can summarise the above findings by considering Bealer's modal argument. Thus, he presents the weaker [i.e., less vulnerable] modal version of his premise as:

> [M]' It is possible that a being could have Pain but lack parts that have 74,985,263 or more functionally related nonconscious parts.[p 371]

Now if this premise is to be employed in a modal argument to show that *this particular pain P at time t* is not the CFS, we must reformulate it as:

> 1(a). It is possible that I could have *this pain* but lack parts that have 74,985,263 or more functionally related nonconscious parts.[p 371]

And as we have argued, it is not at all clear that this premise is true, since there are two kinds of possibility here. If [M]' is to be construed as the intuited modal proposition, we can say that

---

40. Kripke himself suggests that the intuition he has in mind occurs when 'Heat' has a topic-*specific* meaning [e.g.,p 142, final paragraph]. Thus, if 'Heat' were taken topic-*specifically* to mean or refer rigidly to 'Molecular motion', our intuition would indeed be that a sample which is not Molecular motion would not be a sample of Heat. What both he and Bealer appear not to notice, however, is that if, in similar topic-specific fashion, 'this Pain' is understood to mean 'this episode of CFS', or 'this part having at least 74,985,262 functionally related non-conscious parts', we again find the strong intuition that an introspected item which is not one of the latter

although it seems to be true for anyone who has yet to discover [by a posteriori investigation] that it is false [i.e., that in fact the CFS has fewer parts], or for anyone who takes 'Pain' to have a [physically] topic-neutral meaning, the possibility thus conceded can be explained on purely epistemic grounds and therefore fails to convince us that the modal proposition is indeed true. It is an epistemic possibility but not a modal [metaphysical] one. If, on the other hand, it is construed as an expression of metaphysical possibility, we have found no reason to believe that it might be true. To do so would be simply to assume that Pain [construed topic-specifically as CFS], might not have 74,985,263 or more functionally related nonconscious parts; that *CFS* might not have 74,985,263 or more functionally related nonconscious parts. But unless CFS is taken topic-neutrally, to be anything which has properties X, Y and Z, there is no compelling reason for supposing that this is possible.

The same objection can also be levelled, *a fortiori*, against his stronger [more vulnerable] version of the premise, that:

> [M]" It is possible for there to be a being who feels pain but does not have a *multiplicity* of functionally related nonmental parts.

Again, while this might be an epistemic possibility, in order to accept that the metaphysical version of the intuition is present further evidence must be cited. So far, we have found no reason to suppose that the intuition which is present goes beyond an appreciation of the epistemic possibility. For the latter is not merely that we have *yet to determine* that the identity obtains, and

---

would *not* nevertheless be 'this *Pain*'. It appears that both for Pain *and* Heat the intuition evaporates when the term is taken to have a topic-neutral meaning. Kripke fails to acknowledge this in the case of Heat. At least, he offers no convincing evidence for supposing that the intuition in the case of heat would not evaporate under such circumstances, or that the intuition is present in the case of Pain even when 'this Pain' is used topic-specifically.

consequently are able to envisage that it might not turn out to do so. Rather it is the more general point that even if we already knew that Pain is identical with CFS we could intelligibly imagine that it might not have been so. Given just that we have two quite distinct kinds of epistemic access to our physical states, it is intelligible to imagine that what we have epistemic access to in introspection should not be identical with what we have epistemic access to by way of the orthodox scientific route. And clearly we do have two such modes of access. If I place my hand in the fire, I will be able, by normal scientific observations, to determine that my hand has been injured. At the same time, however, it will be possible for me to determine in introspection that I am in Pain, and it seems clear that the Pain is at least a consequence of that damage. Whether the Pain is *identical* with the physical state of damage has not been settled. On the other hand, it is at least intelligible to suppose that it is. And so long as the epistemic situation just described is found plausible we have no good reason to suppose that the intuition of contingency requires any further explanation.

It seems, then, that there is only one other way in which the intuition of contingency might be shown to amount to something more than our epistemic explanation tells us it is. That is, the epistemic explanation must itself be shown to be implausible. We saw earlier that some philosophers might even find it incomprehensible.[41]

Now, in order to make something of this line of defence, the opponent of the identity thesis must do something quite specific. That is, he must be able to cite some characteristic or property which introspectible Pain has but which CFS, or any other physical state with which Pain is supposed to be identical, is at least very unlikely to have. But this suggestion leads us back to the discussion in chapter V. Suppose, for example, that introspectible Pain has the characteristic of being a qualitative experience or quality experienced. The suggestion would then be that it is

---

41. Madell, p 95.

implausible to suppose that any physical state can be a qualitative experience or quality experienced. As we have seen, however, in order to lend any support to this suggestion we would then have to cite some fact which is true of an experienced quality [for example], but which is unlikely to be true of any physical state. We have yet to discover that there is any such fact to be cited, but suppose for now that we find one, and let us refer to this as the fact that experienced qualities have characteristic X. The point now is just that the proposed identity relation is unlikely to obtain just in virtue of our intuitions about characteristic X; which sort of phenomena can have it and which cannot. If we find it implausible that any physical phenomenon can have characteristic X, our intuition to that effect will be just that it seems not to be true. But this is not the *modal* intuition that the proposed identity relation is not necessarily true. It amounts just to the observation that it is difficult or impossible to understand how the proposed identity thesis can be *true*, or that it seems not to be true. Hence, if such a characteristic can be found, the identity thesis will be thrown into doubt just because it seems implausible, or unintelligible. The appeal to the modal intuition that it seems not to be *necessarily* true will have been rendered redundant. For the argument will then take the form:

1. Let Pain be of a particular introspectible type [P].

2. Let C-fibre stimulation be of a particular physical type [CFS].

3. Pain has characteristic X.

4. It is unlikely that CFS has characteristic X.

Therefore,

5. It is unlikely that CFS is identical with Pain.

If this conclusion is taken as expressing an intuition, it can be seen to bear a strong resemblance to the premise [M]2 or [M]3 cited earlier in the present chapter. Thus:

[M]2 The CFS can occur even if it is not P.

[M]3 *It is possible* that the CFS can occur even if it is not P.

Although now, of course, the identity thesis under consideration is of the type-type variety. By the same token, then, it too must be rejected as a candidate for the sought-after modal premise. In order to turn it into a modal premise, we would need to replace 4 with something like:

4(a). It seems that CFS does not *necessarily* have characteristic X.

From which we might hope to infer that:

5(a). It seems that CFS is not *necessarily* identical with P.

But even if the move from 4(a) to 5(a) can be justified by further argument, premise 4(a) still needs to be supported. The question remains, then of what that support might be. The problem in the current context is that we still have no reason to suppose that we have any intuition to the effect that a mental-physical identity relation can only be contingent, other than the intuition which we have already explained away in epistemic terms. Hence, if the modal argument is to have any import, a suitable characteristic X must still be found. We must find an X such that logically priori to ascertaining that CFS does not *in fact* have X, we can say that it seems intuitively that CFS does not *necessarily* have X; that even though we do not know whether CFS has X, CFS *could have occurred* without having X.

Thus, the problem for Kripke here turns out to be parallel to a problem already cited in our discussion of the knowledge argument. For although the intuition seems compelling that qualia are, for example, essentially felt qualities, or essentially possessed by a conscious subject, while physical items are not, that intuition might nevertheless owe its air of compulsion merely to the limited scientific perspective at present available to us. Given further scientific information, it might turn out that the intuition is misguided. Thus, it might turn out that the felt qualities which are possessed by a conscious subject need not be; that although Pain is felt by me it is not *essentially* felt by me. Conversely, although it seems compelling to assume that CFS is not essentially felt by me,

it might turn out to be so.[42] Whether Pain is in fact CFS remains to be established by further a posteriori investigation, by gaining further understanding of the nature and properties of physical items. If Pain and CFS turn out to be identical, then what is essentially true of one is essentially true of the other, since they are but one phenomenon.

Following this line of analysis, then, we find ourselves substantially back at the end of chapter V. Thus, we saw in chapter V that, given the Cartesian premises that:

1. It is of the essence of a phenomenal property that it feel a certain way to its subject [indeed, we might assume that the way it feels is the phenomenal property]

2. It is not of the essence of any physical type that it feel in any way to its subject.

it seemed to follow for MacDonald [p 33] that phenomenal properties cannot be physical properties. Since we are interested in our own occurrent Pains, however, premise 2 will need to be tightened up.[43] We must assume instead that:

---

42. To be fair to Kripke, we should acknowledge that his modal intuition is intended to present a challenge only to 'the usual forms of materialism' [last paragraph of Kripke's "Naming and Necessity", p 155]. It might be legitimate for him to object, therefore, that this latter option, at least, would not amount to a usual form of materialism. Nevertheless, I can find no reason to suppose, at this stage, that it is not an *unusual* form of materialism; a form according to which the type-identity thesis is still true.

43. Using premise 2 would leave open the possibility that all of *our occurrent CFS episodes* do feel in a certain way, even though it is not of the essence of CFS, or even physical types *per se*, that this is so.

2(a). It is of the essence of any physical type that it does *not* feel in any way to its subject.

Hence, if Kripke employs this distinction his position might still be vindicated. But there are two problems. Firstly, if this line of argument is sound, it renders Kripke's entire argument redundant, since now it is evident that we know enough anyway to preclude the identity of phenomenal properties and physical types. Secondly, however, there is no evidence as yet to support 2(a). For in our discussion of Kripke's argument we saw that even if a physical property can be picked out topic-specifically we cannot presuppose that that property is not a phenomenal property. Hence, we cannot presuppose that the physical property does not essentially feel a certain way to its subject, or that Pain essentially feels a certain way to its subject. For if, ex hypothesi, Pain is identical with C-fibre stimulation, the same essential properties will belong to Pain and CFS. As we argued in the previous chapter, there is as yet no reason to suppose that the identity relation should not be discovered as a result of future scientific research.

<u>The Reductive Physicalist's Goal</u>.

Now we might be tempted to concede in defence of Kripke that the above possibility fails to satisfy the objective of the reductive project under consideration. Thus, the possibility we are exploring is of reducing the mental to the physical by discovering that mental phenomena are the ultimate referents of our physical referring expressions. If Pain is rigidly specified as the property or state Rp discernible topic-specifically in introspection, then, it would seem that the identification of a physical phenomenon with Pain could be construed as a reduction in the opposite direction, construing the physical as being of fundamentally mental constitution. What we seem to be saying is that the true nature of, for example, C-fibre stimulation, must turn out on a posteriori investigation to be the phenomenal property Pain. Hence, Kripke's analysis really does succeed in demonstrating that it is not possible to subsume mental phenomena into the existing *physical* ontology. We can see, however, that there are least two reasons why this inference would be unwarranted.

Firstly, if the proposed reduction were successful, the existing physical ontology would not thereby have been *extended* to incorporate phenomenal properties; rather, the items we *took* to be physical items would turn out to be phenomenal properties. We have not yet succeeded in establishing that Pain is anything in addition to CFS, for example, since no characteristic X has been found which Pain, as picked out in introspection, has, but CFS does not. So it remains possible that the 'Pain' we discern in introspection [i.e., possess as conscious subjects] is not *essentially* possessed by a conscious subject, for example, and just is CFS.

Secondly, suppose that science has progressed to the point where all the physical facts about C-fibre stimulation have been determined, and that C-fibre stimulation is seen necessarily to have some of those properties. And suppose, further, that under these conditions it emerges that Pain is identical with C-fibre stimulation. The current suggestion is that under these conditions the *physical* phenomenon will have been reduced to the *mental* and therefore that the objective of reducing the mental to the physical has not been attained. But there is no reason to accept this verdict. All we have established is that the mental and the physical are one and the same; determining the *direction* of the reduction amounts to the further exercise of assigning the fundamental reality, of which Pain/C-fibre stimulation is a member, either to the physical or to the mental realm, or to neither. This is a metaphysical exercise which goes far beyond the mere identification of the two phenomena.

Conclusion.

We have argued that there appears to be no prima facie objection to explaining Kripke's modal intuition, that mental phenomena are not necessarily physical phenomena, in purely epistemic terms. In our discussion of Kripke's argument we saw that even if a token or type of physical state or property can be determinately specified in the scientific epistemic situation EP, it might nevertheless be specifiable only topic-neutrally in an introspective situation Eq. Even if we have a complete command over all the physical facts in EP, then, it still remains epistemically possible that we should be able to pick out physical states or properties only topic-neutrally in Eq. Here we might recall the predicament of Mary emerging for the

first time from her black-and-white room. We found that even with all the physical facts to hand, it would be a further task for her to determine that what she could discern as Rp in introspection was in fact a physical property RP. Determining *that* one is in a particular physical state requires more than the complete information about what that state amounts to. Similarly for Kripke and his Pain, then. Although he might know all about the physical state or property of Pain/CFS, he might be in an epistemic situation Eq in which he does not know that his occurrent Pain at time t is in fact a CFS. For in addition to having a complete command of the physical account of Pain/CFS, he would need to learn to recognise introspectively that he is in that state. Similarly, he might be able to determine in Eq that each of the occurrent pains he has is of the type [Pain], but still not have sufficient information to determine that Pain is CFS. And this is enough to explain why the identity of both the *types* [Pain]/[CFS] and any *tokens* of Pain/CFS can seem to be only contingent. For it seems that either identity relation might not obtain, or might not have obtained.[44]

So it now seems that the only way of upholding Kripke's refutation of an identity relation between Pain and C-fibre stimulation is by an appeal to the intrinsically implausible nature of such a relation. That is to say, we must show that there are independent and compelling reasons to doubt that it could turn out to obtain. Thus, we might try to show that our introspected Pains have some characteristic X which no physical properties are likely to have. Having explored the principal candidates for characteristic X in chapters IV and V, however, it is clear that there is further work to be done.

---

44. To render this possibility more plausible, we might again imagine a humanoid robot which has a complete command of all the physical facts about Pain/CFS. Even then, it would require further programming in order to determine introspectively *which physical state it was in* at any particular time. This seems so obvious as not to merit further demonstration. If human beings are purely physical beings, then, a similar situation obtains. Hence, we cannot presuppose that the introspected phenomenal properties are not just physical properties.

We are not entitled simply to *assume* that introspected properties are epistemically private, or that they are essentially possessed by conscious subjects, for example, since neither of these characteristics has yet been shown to belong to such properties. To assume that they are would at this stage be to *presuppose* that those properties are not physical properties belonging to [S]. To argue in modal style that Pain *seems* to be *necessarily* possessed by a conscious subject, or epistemically private, would then be entirely pointless. For in view of the epistemic considerations already explored, we could explain this apparent necessity as follows. It seems that Pain is essentially or necessarily possessed by a conscious subject just because we are presupposing that Pain is not CFS. Similarly, it seems necessary that Pain is epistemically private just because we are presupposing that pain is not CFS.

Nevertheless, there is a further intuition to be addressed by the reductive physicalist. For in spite of all the arguments offered so far, the dualist might claim that it is just obvious that there is *something* about introspected properties which is not true of physical properties. Thus, if a sensation of pain really can be shown to be more than just a physical property discerned in introspection, there might be good reason for inferring that sensations of pain are not physical, in the sense outlined in our introduction. For it might then be plausible to suppose that sensations of pain really are epistemically private, or essentially possessed by a conscious subject. More generally, it might be plausible to suppose that introspectible pains, sensations of red, and so on have, or are, qualitative characters or "felt qualities" which a third-person physicalism simply fails to take into account. It seems that it is extraordinarily difficult to formulate an intuition of this kind in a convincing way. One reason why this is so, I believe, is that we have yet to understand the full implications of a third-person physicalistic account; the proposal that all occurrent states and properties are fully intelligible and epistemically available in the third-person perspective. Irrespective of how this question might be explored further, however, it seems clear that Kripke's approach, like those already examined, can only succeed on the *presupposition* that sensations are as this intuition suggests they are.

THE PROPERTY DUALISM ARGUMENT

Introduction

We saw in the previous chapter that Kripke's argument for the non-identity of phenomenal and physical properties suffers from two outstanding difficulties. Firstly, the argument simply assumes that there are phenomenal properties to which it is possible to gain direct access by introspection. We have not belaboured the difficulties invoked by this assumption, however, since even if it were true the argument suffers from a second, and fatal flaw.

The second is the epistemic point that, even if the first assumption is true, it appears that the paradigmatically physical properties or states with which introspected phenomena are supposed to be identical cannot be determinately picked out as those properties or states *a priori*. By this we mean just that there are occurrent epistemic states Eq in which it is possible to discern phenomena introspectively and yet not have sufficient information to establish that those phenomena are those physical properties or states. Having identified a pain in introspection as such, for example, it remains to be established by further investigation whether that pain is a physical property or state. Hence, there is a prima facie objection to Kripke's inference that phenomenal and physical properties are distinct. We have, as yet, no prima facie grounds for presupposing that science should fail to establish that this is so. Even if science has reached the point where the two phenomena are known to be identical, however, it still remains possible for Smith to be in an epistemic situation in which he knows that he has a pain but does not also know that it is identical with a physical state or property P. All that is required for this epistemic situation to obtain is that in Eq Smith is able to discern his [physical] pains only topic-neutrally; just as a robot which has a complete command of all the physical facts and is in physical state P might be able to register the fact that it is in a state of type [p] without also being able to determine even that state p is any physical state at all. For Smith in that situation, then, his complete understanding of the physical account of discerning p in introspection will fail to furnish him with the ability to determine *in introspection* that he

is in state P. Our reductive physicalism tells us that state p *is* state P, but it does not entail that being in state P is sufficient to determine that one is in state P, or even that being able to determine that one is in state p entails being able to determine that one is in state P.

The problem for Kripke, then, is that his intuition that p is not *necessarily* P can be explained away on epistemic grounds. It can be explained as the ability to understand the proposition that being in state p, as picked out topic-neutrally in introspection just as p, might not have amounted to being in state P. It is, in effect, just a tacit recognition of the epistemic fact that phenomena are discerned in introspection only topic-neutrally with respect to which physical phenomena they might be. The obvious recourse here for Kripke is to some further characteristic of introspected states or properties in virtue of which our topic-neutral account can be shown to be incomplete. Thus, if he could establish that p is something discerned topic-*specifically* in introspection he might then appeal to the intuition that p seems not to be necessarily identical with P. We have as yet been unable to find any state or property which satisfies this description. Even if we had, however, it would seem that the modal intuition would be redundant. For if it could be shown that p is picked out topic-specifically in introspection and yet not be known to be P, it would be tempting to infer straight off that p and P are distinct. Our reductive physicalism requires that all states and properties discernible in introspection are also fully accountable in the third-person perspective, and hence that *any* topic-specific recognition of those states and properties would enable us to recognise them *as those properties.*[45]

---

45. Even then, there is, at least in principle, another possible line of argument which would present a challenge to Kripke's argument. Thus, it might be suggested that physical properties or states, as picked out in paradigmatically scientific fashion, are specifiable only topic-neutrally, and hence might still be identical with intropsected states or properties. Bertrand Russell [e.g., 1921, 1927] developed his "neutral monism" along these lines, and more recently John Foster [1982] and Michael Lockwood [1981 , 1989]

If we take it that our topic-neutral account of p is complete, then, we shall need some other way of challenging reductive physicalism, and the property dualism argument promises to do just that. Thus, instead of setting out to show that the p is distinct from P, we might assume for now that they are indeed identical. But Stephen White argues that even then the *properties* via which a physical state or property is picked out epistemically must be different in the paradigmatically scientific and introspective perspectives respectively. White's aim is to establish that although mental referring expressions refer topic-*neutrally* to neural properties or states; they refer topic-*specifically* to dispositional or functional characteristics which are grounded in those neural referents. Since this would entail that the dispositional characteristics through which mental referring expressions refer can be known *a priori*, however, and we have already dismissed this claim in chapter IV as being impossibly difficult to substantiate, however, his project would seem to be doomed from the outset. What remains possible is that his argument might be adapted to show just that the a posteriori coreferentiality of mental and physical referring expressions is symptomatic of a fundamental difference between

---

have made impressive attempts to develop the theme. We do not explore this approach here, however, since our own objections would seem sufficient in the context of the present discussion.

I take Trenton Merricks [1994] to be offering a *metaphysical* [modal] version of this sort of argument. Thus, translated into the terms of our own discussion, his point might be that we could accept Kripke's modal proposition, that pain is not necessarily a *physical* property, without having to give up the Pain/CFF identity thesis. And he suggests that we could do so by conceding just that CFF is not *necessarily* physical. But Kripke does not *need* CFF to be necessarily physical. Thus, even if "CFF" refers only topic-neutrally to some *physical* state or property P, and CFF is therefore only contingently physical, the Pain/CFF identity thesis still entails that Pain is *necessarily* identical with CFF. In order to evade this entailment, we need to show that "Pain" refers only topic-neutrally to CFF. It is this suggestion that we found to be plausible at least *prima facie* in the previous chapter.

mental and physical properties. If it can, we might use White's reasoning to infer that there are two fundamentally distinct property types; phenomenal, or mental, and physical.

One final point must be cleared up at the outset, concerning White's use of the notion of an identity relation which can be known a priori, or a posteriori. When he says that an identity relation can be known only a posteriori, he should be understood to be making an epistemic point about *first-person* knowability. Thus, when he says that the referent of "the evening star" can be known only a posteriori to be identical with the referent of "the morning star", he must be taken to be saying that this epistemic situation obtains *for the user of the two expressions.* His delivery is apt to be misleading on this point; he cites "Smith's pain at t", for example, as if it can be known by a third party to be coreferential with some other expression only a posteriori. Clearly, the third person interpretation would be trivially true for any pair of referring expressions which co-refer. In order for Jones to establish that any two of Smith's expressions co-refer at all it would be necessary for Jones to interpret Smith's utterances on a particular occasion, and such interpretation would itself constitute an a posteriori investigation. What White needs to establish is that Smith himself is unable to establish the co-referentiality of his own expressions a priori.

What, then, would it amount to in White's argument for an identity relation to be knowable in the first-person *a priori*? According to our previous considerations, we might suppose it to entail being able to infer logically from the fact that one is experiencing a phenomenal property p that one is experiencing a physical property P, and that the two are identical, without recourse to any additional information. Hence, the fact that p is being experienced entails logically that P is being experienced and that p is P. In short, we might say that in consequence of our epistemic explanation for Kripke's intuition, p and P would have to be logically or conceptually identical for their identity to be known a priori; p would have to be discerned topic-specifically as P in introspection. The position for White, however, is that p is indeed discerned only topic-neutrally in introspection, and can therefore be known to be P only a posteriori. Thus, he accedes to our own claim that even if the states Smith is experiencing and refers to as "my headache p at

t" and "my C-fibre stimulation at t" were identical, he cannot know *a priori* that the two expressions co-refer. Hence, we might surmise that for White it is at least not a logical or conceptual truth that they do so. If we assume that Smith can at least know that he is in a state of the type [headache], then, he is unable to infer logically or conceptually from this fact that the state is of any particular physical type [P], even if it is. Thus, epistemically, the type [headache] is topic-neutral with respect to any paradigmatically physical type [P]. Here, we are reminded of the epistemic state Eq in chapter III which Smith can be in with respect to his headache. But White takes this as his starting point in order to offer a new counterargument to reductive physicalism. Thus, instead of inferring that "My headache p at t" must not corefer with "My C-fibre stimulation at t", as Kripke does, he infers rather that even if they *do* corefer [to brain-state X], our referential route to brain-state X is via epistemic *modes of presentation* of two distinct types.

As we shall see later, this interpretation seems justified since White's desired inference is that epistemically we identify the referent of each expression respectively via the recognition of *properties* of different types, one paradigmatically physical and the other physically topic-neutral. Barring the topic-neutrality of *physical* referring expressions, he could only have any hope of justifying that inference if the type [headache] were assumed to be epistemically topic-neutral with respect to any paradigmatically physical type.

The Argument Presented by White.

The property dualism argument itself is succinctly encapsulated in Stephen White's *The Curse of the Qualia* [White, 1986].

The general principle is that if two expressions refer to the same object and this fact cannot be established a priori, they do so in virtue of different routes to the referent provided by different modes of presentation of that referent. These modes of presentation of the object fall on the object's side of the language/world dichotomy. In other words they are aspects of the object in virtue of which our conceptual apparatus picks the object out; they are not

aspects of that conceptual apparatus itself. Hence the natural candidates for these modes of presentation are properties. ... Since there is no physicalistic description one could plausibly suppose is coreferential a priori with an expression like 'Smith's pain at T', no physical property of a pain (i.e., a brain state of type X) could provide the route by which it was picked out by such an expression....

> This argument, which I shall call the property dualism argument, shows that unless there are topic neutral expressions with which mentalistic descriptions of particular pains are coreferential a priori, we are forced to admit the existence of mental properties. [Stephen White, 1986, pp 92-3].

In order to evaluate White's position here we should perhaps begin by trying to distil out the logical structure of his argument. The following would seem to exhibit the most natural interpretation. Here, we adopt White's assumption that the following argument would be sound *unless* there were topic neutral [dispositional] properties of the sort he seeks.

> Premise 1. For any two coreferring expressions A and B which are not knowable *a priori* to be coreferential, the mode of presentation associated with the referent of A must be logically distinct from the mode of presentation associated with the referent of B.
>
> Premise 2. Modes of presentation are properties of the entities they present.

Taking an example in which one of the referring expressions is physical and the other mental, then, the conclusion is that;

> 3. If "Smith's pain at t" and "Smith's C-fibre stimulation at t" (or whatever physiological expression refers to Smith's pain state) are not knowable *a priori* to be coreferential, then their respective modes of presentation are logically distinct properties.

White's supporting considerations run as follows, apparently in the form of a somewhat veiled *reductio ad absurdum* argument.

Suppose that this is not the case. Suppose, that is, that two descriptions are coreferential and that [in the first person] this fact cannot be established a priori and has not been established a posteriori. *And suppose that there are not two different properties in virtue of which the two descriptions pick out the same referent.* That the descriptions are not coreferential a priori (and not known to be, a posteriori) means that there is a possible world in which speakers who are epistemically equivalent to us use these terms to refer to different objects. There is, for example, a possible world in which the inhabitants are epistemically equivalent to those of our ancestors who used "the morning star" and "the evening star" before the discovery that the terms were coreferential and in which the inhabitants use the terms to refer to different planets. As used by the inhabitants of this possible world, these terms must pick out their referents in virtue of distinct properties because, unlike our terms, theirs pick out different objects. Hence the expressions as used by our ancestors must, contrary to our assumption, pick out their common referent in virtue of two logically distinct properties of that referent." [p 92]. [My emphasis and first parentheses]

1

The Structure of White's Argument.

Now while it seems clear how the above defence is intended to run, there is some confusion as to which of the numerous assumptions are actually instrumental in the demonstration, and which are merely redundant. It appears that we could reduce the defence to the following form. Taking the familiar example of the Morning star and the Evening star, we can imagine an earlier time before the two expressions had been found to be coreferential. At that time, and given the limited information then available, it was still logically possible [from the user's point of view] that they should turn out not to be coreferential. Upon further astronomical investigation it might have turned out [in White's other possible world] that "The Morning star" referred to one planet and "The Evening star" referred to another. But this could only have been logically possible at that time if the respective referents of the two expressions were picked out in virtue of two logically distinct properties.

This seems to be essentially the argument White is employing, but if that is the case we can see that the assumption he makes that "there are not two different properties in virtue of which the two descriptions pick out the same referent" is simply redundant. At no stage in the argument is this assumption employed to develop the reductio. Certainly, as he points out, his final conclusion is contrary to that assumption, but the said assumption has not been employed in any deductive process of reasoning in the course of the argument. We are misled into expecting a reductio demonstration by the inclusion of this redundant assumption. What he actually appears to be arguing is that since, in any case where the [factual] coreferentiality of two expressions cannot been established a priori [we shall refer to this from now on as the *"APC" condition*], it is logically possible that they should turn out to be either coreferential or not, we must be picking out the referent in each case by way of two distinct properties. Construed thus, White's further argument for property dualism in the Fregean example turns out to be a straightforward reiteration of his original demonstration.

There is, however, another respect in which White's second version of the argument might be seen as representing some improvement over his first. Thus, whereas in the standard Fregean example of "the morning star" and "the evening star" there might be some doubt as to whether or not the two expressions co-refer by way of different properties, White implies that the same doubt cannot be cast over the other-worldly example in which the two expressions do not even co-refer. For if, in the latter case, the two expressions in fact refer to *different objects* it follows for White that they cannot possibly so refer via one and the same property. If we accept this conclusion the rest of the argument can proceed along the following lines. Our ancestors and those of White's other world are in an epistemically identical position. For both, it is logically possible that "the morning star" and "the evening star" should turn out to refer either to one and the same object or to two distinct objects. But since we have already accepted that in the case in which the expressions do *not* turn out to co-refer they must refer via logically distinct properties, we must now accept also that in the epistemically identical case of our own ancestors a similar situation obtains. They too must be referring, albeit in this case to one and the same object, via two logically distinct properties.

Indeed, if our earlier epistemic considerations were sound, we might offer a bolder version of White's claim. Thus, we can say that even if Smith *already knows* that the two referring expressions refer to the same object, the fact that they do is knowable only a posteriori in the first-person. For in view of the topic-neutrality of the referring expressions, he can still say that it is *epistemically* possible to know that one is discerning the morning star without also knowing that it is the evening star. The latter fact was discovered *a posteriori* by learning that the morning star *is* the evening star.

We can now begin to see how White's argument might be employed for our own purposes. Firstly, the structure of the argument might be outlined in the following way. Firstly, we *assume* that a mental referring expression corefers with some physical referring expression A. The argument is that even if this is true, the properties through which each expression refers must be of different types. Thus, on that assumption:

1. Co-referring expressions which refer via token properties of a single type can be known to be coreferential a priori [i.e., logically or conceptually].

2. A mental referring expression M cannot be known to be coreferential with *any physical referring expression A whatever* a priori. [This will be referred to hereafter as the "*Universal* APC" or "uAPC" condition].

Therefore,

3. M and any A whatever refer via token properties of different types.

4. For any physical property p whatever, there is some A which refers via p.

5. Every referring expression refers via a property.

Therefore,

6. M refers via a non-physical property.

Premise 2 reflects the premise employed by White in his own argument, but we can see that by making it more specific some of the ensuing premises might be rendered redundant. Thus, we might replace 2 with:

2. A mental referring expression M cannot be known to be coreferential with any physical referring expression A *with which it might be coreferential* a priori.

And in line with this narrower premise we might then legitimately construe inference 3, and premises 4 and 5, to apply just to the candidates for A specified in 2. This is a relatively unimportant refinement for our purposes, however, as we shall see.

Once this conclusion has been reached, the question arises as to what sort of property M refers through. There are two possibilities. Either it refers to the referent R via an irreducibly *mental* property, or it refers to R via a property which is physically grounded but *topic neutral* with respect to the physical type which R belongs to. For White, the positing of mental properties is absurd, and on that assumption he is able to infer that the topic-neutral alternative is the correct one; a physically grounded property of a dispositional or functional character provides the route to the referent R. But this would entail that epistemically our mental referring expressions could be known a priori as referring via dispositionally characterised properties; a result which we rejected as being just too difficult to substantiate in chapter IV. Hence, if the argument outlined in 1 - 6 is sound, we have no prima facie reason to reject the alternative that irreducibly mental properties are involved.

White's Argument as a Supplement to Kripke's.

Thus construed, White's argument in 1 - 6 can be seen as a natural development from the argument presented by Kripke, as interpreted *epistemically* in the previous chapter. For according to that interpretation, it would be true to say that mental phenomena in particular are designated only topic-neutrally in the first person perspective, and therefore that we seem to have no a priori grounds for precluding the identity of C-fibre stimulation with pain. But

White's argument can now be employed as a further attempt to establish that there is something intrinsically dissonant for physicalism about a mental-physical identity relation which is not knowable by the bearer a priori. Assuming this APC relation to obtain, let brain state R at t be the fundamental referent of the two expressions "Smith's pain at t" and "Smith's C-fibre stimulation at t". If their respective referents cannot be known by Smith *a priori* to be identical, says White, it follows that the referent is being picked out in each case via a different property. As we have seen, and contrary to White's expectation, the only plausible conclusion would be that the mental expression refers via an irreducibly mental property.

In effect, then, we are now sidestepping Kripke's unsuccessful appeal to a metaphysical intuition. If the two expressions are coreferential, the APC condition entails that one fundamental referent exhibits both mental and physical properties. White rejects this conclusion on metaphysical rather than epistemic considerations; he finds the mental-physical property dualism implied by the epistemic considerations intrinsically unacceptable per se. What concerns us here, however, is just whether the argument succeeds in showing that property dualism *of one sort or another* is the inevitable consequence of the APC condition, as the argument 1-6 purports to establish. The question of whether a mental-physical property dualism is entailed in certain circumstances is a further issue which need not be addressed at this stage.

Although White's argument is initially compelling, there are two further assumptions hidden within it which deserve further scrutiny. Firstly, is it really true that, whenever the APC condition applies to two *physical* referring expressions, they must be referring to the common referent via logically distinct properties? Secondly, even if they must, how can White justify the further assumption that in uAPC cases involving a *mental* referring expression the latter must be referring via a property which is not just a paradigmatically physical property? We can now consider these questions in turn.

1. The APC of Physical Referents.

The structure of White's argument is such that premise 1 of the argument must be substantiated in the first instance. Premise 1 clearly entails that there are no pairs of referring expressions to which the APC condition applies and yet the two expressions refer via token properties of a single type. If we can establish that this is false, then, the argument as presented cannot even get started. In the discussion which follows, an attempt will be made both to expose a fundamental flaw in White's argument and to show that in any revised form the property dualism argument must fail. The objection which we will attempt to substantiate is that on any interpretation of "logically distinct properties" which would imbue White's conclusion with the force he assigns to it that same interpretation renders his premise 1 false. In other words, the argument can only go through on an *equivocation* over the meaning of "logically distinct properties". As we shall see, the problem is that no clear distinction between *token* and *type* property differences is maintained.

The equivocation can be brought out initially by reference to the Fregean example. In that example, the APC condition clearly applies to the two expressions "the morning star" and "the evening star". It is equally clear that the reason for this is that the two terms co-refer to Venus by virtue of distinct modes of presentation. The fact that the two distinct modes of presentation share a common object must be established a posteriori. Nor does there seem to be any difficulty, in this example, in regarding the two modes of presentation as entailing properties of distinct types. The first is the property of being the last star to be visible in the morning and the second is the property of being the first star to be visible in the evening. The two presentations are epistemically dissimilar and it is therefore simply a matter of a posteriori fact [for the user] that the two expressions turn out to be coreferential. In terms of topic neutrality, we can see that even the *type* [morning star] is epistemically topic neutral with respect to the type [evening star]. The first signs of equivocation emerge, however, when White attempts to strengthen his case by appeal to the other-worldly example. For while numerically distinct planets must be identified epistemically via numerically distinct *token* properties, the example offers no

additional evidence that the properties are of distinct *types*. Changing the example slightly will help to clarify this point.

Thus, suppose that in another possible world there are *two* morning stars; two distinct planets either of which appears randomly, from one morning to another, as the last star visible in the morning. Since the inhabitants of this world are unable to distinguish between the two, then, we can say that reference to each proceeds via epistemically indistinguishable properties. That is, the properties are of the same *type* even though the two planets exhibit numerically distinct *tokens* of that property. But then the fact that there are two planets singularly fails to indicate that different property types are involved. If the properties are of different types, then, they must be shown to be so even if the referent is a single planet. The observation that there might have been two distinct planets offers no additional support for this claim. More seriously for White, it seems that our example exposes his premise 1 as being simply false. For while the inhabitants of the other world cannot know a priori that the expression "the morning star" refers to one and the same planet on each occasion, it is nevertheless a fact that on each occasion it refers epistemically via token properties of a single type.

Turning to the case of Smith's brain state R at t, we can see that parallel considerations apply. Since the two expressions "Smith's pain at t" and "Smith's C-fibre stimulation at t" cannot be established a priori to be coreferential they must each refer to the supposed common referent R by virtue of distinct *token* modes of presentation, or properties, as White would have it. But further argument is then required to demonstrate that these distinct property tokens belong to interestingly different *types* [one paradigmatically physical and the other not]. Whether this demonstration can be made to work remains to be seen. Our initial task is to find out whether, in such a case, a type-type property dualism of *any* sort can be inferred. To this end, we might begin by trying to approach the problem on purely physicalistic premises. Thus, we might begin by replacing the person Smith with an entirely physical robot, or zombie, which is capable of collecting and assimilating all the information needed to establish both [although perhaps on different occasions] that it is in pain and that its C-fibres are being stimulated.

A Purely Physical Counterinstance to White's Premise 1.

Suppose that the robot conducts a physiological examination of its own internal state R by inserting a probe into its head and taking readings from an external instrument, which we shall name the "fibroscope". Since it has two eyes, the robot is clearly able to glean the required information via one eye or the other. On the assumption that the physiological [or electronic] processing of the information through each eye is substantially of the same type, we can then say that the mode of presentation of the robot's physical state R is of the same type in each case. But we have no prima facie reason to assume that the robot is able to determine *a priori* that the state detected via each eye is numerically identical. Further internal circuitry would be needed to provide it with that information. Thus, the robot would have to be wired in such a way as to "know" that the information gathered through each eye refers to one and the same physiological state [it might even be unable to determine that the fibroscope viewed by each eye in turn is one and the same instrument]. There is no compelling reason to suppose even that the robot should be able to recognise that the physiological state as detected via each eye is of one and the same type. Even if there were, however, the robot could not be said to *know* that the identity obtained without having access to some way of checking and validating the information provided by its own circuitry. But this is a paradigmatic case of a posteriori coreference. What we mean by "a posteriori" in this context is just that there are possible epistemic situations for the robot in which its knowledge that the referent to which it has access via one eye is identical with the referent to which it has access via the other. Hence, there might be insufficient information available, in a particular epistemic state, to justify the logical inference that the identity relation obtains. Thus, while the state R detected via each eye is presented to the robot via distinct *token* properties, the APC condition obtains even though the token properties are of the same *type*. And this shows again that White's premise 1 is not a general truth.

A similar observation can be made about Smith. For while he might be in an epistemic situation in which he is able to determine that he is in some brain state R just by reading the fibroscope with his left eye, and some brain state R' just by reading the fibroscope with his right eye, he might require further information to

determine that the same state, or even the same fibroscope, is being detected via each eye. On the assumption that he has some understanding of his neurophysiological make-up, and has learned that each eye is pointed in roughly the same direction in physical space, he is likely to infer that one and the same state is being detected in each case. But that assumption invokes information which constitutes a posteriori knowledge about himself and the external world; information which might not be available in a particular epistemic situation in which Smith can nevertheless determine via each eye respectively that he is in R and R'. Hence, the APC condition can obtain for Smith even though the unique state detected by each eye is presented to him epistemically via modes or properties of a single *type*.

The physical counterinstance just cited was directed specifically at premise 1 of the property dualism argument as applied to Smith's epistemic situation. Thus, even if the referring expression "C-fibre stimulation", as used by Smith when reading the fibroscope with the left and right eye respectively, is known by science to co-refer, it remains possible for a Smith who lacks the relevant scientific information not to be aware of that identity relation. In order to discover that the relation obtains, he would need to acquire further scientific information. And the example shows that even in this situation the token properties through which the expressions refer might not be of distinct types. We might even be entitled to say that, if the two references are topic-neutral, even Smith's *complete* physical grasp of the nature of both R and R' *per se* would leave him needing further information to determine that the referents discerned topic-neutrally in the first-person perspective are one and the same. For as we have argued previously, knowing all the physical facts about the referents *per se* might not be sufficient for knowing topic-specifically that one is discerning those referents on a particular occasion [as with Mary emerging from her room for the first time].

A Mental-Physical Counterinstance to White's Argument.

Since premise 1 is evidently indefensible as a general thesis, it will have to be replaced with a more specific premise which nevertheless enables White to draw the required conclusion. Thus, we

might suggest that at least in cases involving a *mental* referring expression, premise 1 is true. But this would be unnecessarily restrictive. All we need to establish is that in such cases there are properties of two different types, one mental and the other physical. So the argument might then be reformulated as follows: Again, we assume that M *is* coreferential with some physical referring expression A.

1'. Co-referring expressions, one of which is mental [M], and the other physical [A], and such that [M] does not refer via any *non-physical* property, can be known to be coreferential a priori [i.e., logically or conceptually].

2. A mental referring expression M cannot be known to be coreferential with any physical referring expression A whatever a priori [uAPC].

Therefore,

3'. M *does* refer via a non-physical property.

In this argument, our 'new premise 1' enables us to eliminate the remainder of the previous argument. For if 2 is to be employed as the uAPC premise, 1' will be sufficient for our purposes. Thus, if 2 is true, the desired conclusion in 3 will follow just if every M which does *not* refer via any non-physical property can be known a priori to corefer with some A. There is no prima facie reason for supposing that 1' might be rendered any more plausible than the more general premise 1. For if the two referring expressions are of significantly different types, one mental and the other any physical property, or even no property at all, it seems *less* likely that they should be known a priori to co-refer, even if they do so without recourse to any mental properties. But since premise 1' is appropriate, the argument is still valid. In order to show that it is unsound, then, we might now look for a counterinstance in which premise 1' is false.

In order to avoid *presupposing* that the human subject is not a purely physical being, we are entitled to assume at this stage that a purely physical robot, or zombie, might be *physically and dispositionally indistinguishable* from the human subject [and might

indeed turn out to *be* a human subject]. Thus, we might imagine that, like us, the robot is able to make the discovery that it is in pain [computer-state R] by way of an entirely internal [or "mental"] route, without recourse to any form of "physiological" examination [in the sense that a neurophysiologist might apply scientific tests to establish that Smith's C-fibres are firing], and that it uses the expression "this unit's pain at t" to refer to R via this internal route. We can also imagine that it is a sufficiently accurate facsimile of a human being also to have the ability to carry out such "physiological" examinations on itself, and uses the expression "this unit's C-fibre stimulation at t" to refer to R via the physiological route. So far, then, we are entitled to assume that a purely physical robot might have epistemic modes of access, of two distinct types, to a single physical state [computer-state R at t]. We are reminded here of the epistemic state Eq in chapter III. In that state, Smith was able to determine that he had a headache without even knowing that it was a physical state of any type whatever.

In this example we can concede, in deference to White, that although one of the referring expressions is indeed mental and the other physical, the modes of presentation envisaged for our robot are now of significantly *different types*. In the one case the robot learns that it is in the relevant state by way of its internal circuitry. Electrical stimulation of the C-fibres leads internally to the stimulation of its "judgement centre", where the judgement that "this unit is in pain" is thereupon deemed to be true. This is the robot equivalent of the process by means of which Smith is able to determine introspectively that he is in pain. In the other case it learns by way of an external examination of its own physiological state, with the help of the appropriate scientific instruments, that the judgement "this unit's C-fibres are firing" is true, and refers to the state R it is then in as "this unit's C-fibre stimulation at T". This second route to the referent, then, corresponds to the route by which Smith, or for that matter anyone else, might determine by neurophysiological means that Smith is in pain.

So the concession to White must be that the properties involved in this example are of *different types*. But the important point is that they are both purely *physical types*, and therefore that it represents a genuine counterinstance to premise 1'.

There is, however, an immediate objection to this line of approach. The robot, it might be insisted, is too artificially contrived an example to bear much relevance to the real case of human pain.[46] In particular, we began by making the assumption that the robot is a purely physical being which is physically and dispositionally indistinguishable from ourselves. Surely, the objection would run, this begs the very question we are trying to answer; namely, whether a purely physical being *could* satisfy that description. The problem with this approach is that we cannot even be sure that such a robot is logically possible. Thus, if it is dispositionally indistinguishable from us, can it also be *physically* indistinguishable from us, or even purely physical in constitution? White's thesis refers specifically to the *human* condition. If our example is to constitute a genuine problem for White's argument, then, the robot must, strictly, be purely physical and constituted just as we are. But then we cannot be sure that the robot would even be equipped to make the judgement "this unit is in pain". To suppose that it would be so equipped would amount to *presupposing* that the property via which the expression refers to its physiological pain state in the human case is a purely physical property; precisely the point in question. In order to evaluate this objection, then, we need to consider in more detail exactly what the robot example shows, and what it does not show.

What the example does not show is that, given the physiological make-up of a human subject, Smith would be capable of introspecting his own pain state in the absence of non-physical properties [i.e., that the absent qualia possibility obtains]. We have as yet insufficient knowledge of human physiology to come to a decision on this matter. Hence, we are not entitled to assume that the envisaged

---

46. Jeff McConnell [1995, p 181], for example, responds in this way to Brian Loar's argument [1990, pp 84, 87-8] that sensory discrimination need not be assumed to amount to anything more than a *recognitional disposition*, without the intervening phenomenal properties. Although McConnell might turn out to be right about this, we have yet to find a compelling reason to suppose that he is. What we are entitled to assume, however, is that we have yet to discover whether such a robot is possible.

possibility of such a robot would undermine White's argument. Hence, because the robot is physico-dispositionally exactly like the human subject, he is not entitled to *assume* that no such possibility exists. As it stands, then, our version of White's argument is unable to provide the conclusion that the qualia-dualist requires. The property dualism argument would only be sustainable if a purely physical simulacrum of ourselves were not possible, and this is the point at issue. And since we do not yet know that such a robot is impossible we are unable to infer from White's argument that anything of a non-physical nature is occurring in our own case.

A second objection to our counterinstance might be that premise 2 is too general. Thus, instead of having to claim that:

> 2. A mental referring expression M cannot be known to be coreferential with any physical referring expression A whatever a priori [uAPC].

it would be sufficient for the purpose of the argument to claim just that:

> 2'. A mental referring expression M cannot be known a priori to be coreferential with any physical referring expression A *with which M might plausibly be supposed to be coreferential.*

It is quite obvious, however, that this refinement would be of no avail. For if we had independent reasons for claiming that there is no plausible candidate A whatever, the entire argument would be rendered redundant, since it would follow immediately that the proposed identity relation between mental and physical referents is implausible *per se*. In order to avoid begging the question, then, we must allow at this stage that there is at least some plausible candidate A to be considered in the human case. Thus, for example, we might assume that "my pain at t" and "my C-fibre stimulation at t" are at least plausible candidates for coreferentiality. And we have argued at length that the coreferentiality of such expressions in the human case can only be known a posteriori, if at all. Hence, in the human case, we are entitled to assume that there is an M and an A which comply with the requirements set out even in 2'.

And again, we are not entitled to *presuppose* that the counterexample could not obtain for a purely physical being which is physico-dispositionally just like us. Since premise 2' is acceptable in the human case, therefore, we are entitled to assume for the sake of the argument that premise 2' could also be true for the robot. Hence, even if premise 2' is adopted, we are entitled to assume that it *might* be true even if we are purely physical beings. And this again implies that our version of the property dualism argument can only operate successfully on the *presupposition* that human beings are not purely physical beings; again, precisely the point at issue.

White's Topic-Neutral Alternative.

It would be legitimate for White to point out at this stage that he is not actually arguing for the occurrence of mental properties. What he is saying, rather, is that in cases where the uAPC condition [or premise 2'] obtains the mental referring expression must refer either via a mental property or via a [neurally] topic-neutral dispositional characteristic. Since he finds the former explanation absurd, he will claim that he is entitled to infer the latter. Now we have seen that the property dualism argument per se fails to indicate the occurrence of any properties which might be regarded as non-paradigmatically physical at all, and therefore that as it stands it cannot be employed for his purposes. In any case, we saw earlier, in chapter IV, that there seems to be no hope of establishing that *dispositionally* or functionally characterised physical types bear an a priori identity relation with phenomenal types, and this conclusion effectively ruled out his desired position. Nevertheless, an *a posteriori* identity relation between phenomenal and dispositionally characterised physical referents remains possible. But since the uAPC condition [or premise 2'] in general fails to indicate the occurrence of any non-paradigmatically physical types at all, it can hardly be expected to indicate that a mental referring expression must refer via a topic-neutral dispositionally characterised property.

The general position with regard to White's argument can now be stated in the following way. Firstly, it must be borne in mind that we are attempting to employ the property dualism argument to establish that there are properties of two distinct types, one

physical and the other mental. White thinks that *if* there were only properties of these two types available, properties of these two types would be implicated in cases involving a mental referring expression and an expression which refers explicitly to a physical state of a paradigmatically neurophysiological type. He tries to avoid the conclusion that there are mental properties, however, by offering the alternative inference that the mental referring expression refers a priori to a physical state characterised, topic-neutrally, in terms of a dispositional or functional type [D]. Since we have shown the latter inference to be unavailable, then White's argument, if successful, *would* entail the existence of mental properties. Similarly, and contrary to White's position, if it were successful we might infer the existence of mental properties when the physical referring expression refers explicitly to a topic-neutral, physically grounded but dispositionally characterised state. But the argument is *not* successful as it stands. Hence, in general, we can say that in cases of uAPC involving one mental referring expression and any *physico-dispositional* referring expression whatever the existence of mental properties is not entailed. In order to establish that entailment, then, further facts about mental properties in particular would have to be invoked.

The Attempt to Supplement the Property Dualism Thesis.

In order even to get started with this demonstration, we need to be equipped with a satisfactory demarcation between mental and physical properties at the outset. Now, there are two ways in which we might demonstrate that properties fall into two irreducibly distinct categories. One way would be by simply *defining* the set of mental properties as comprising just those properties which are epistemically related to all physical properties in the way White implies. Thus, on the assumption that the uAPC relation obtains between a mental referring expression and *any physico-dispositional referring expression whatever*, we might suggest that those expressions must refer via a mental and a physico-dispositional property respectively. The problem with this approach, however, is that the proposed definition of mental properties would not be sufficiently selective. For, as we saw in the case of Frege's example and of the humanoid robot, pairs of referring expressions which even White would regard as being purely physical can be found

whose referents are knowable only a posteriori. Any attempt to define mental properties, or mental referring expressions, along these lines would therefore need to be supplemented with some further criterion by which to demarcate the mental from the physical.

The problem for the qualia-dualist, then, is to find something interesting to say about mental properties in general. So let us assume, for the sake of argument, that he has indeed found some characteristic of all mental properties which serves to demarcate them from physico-dispositional properties. As an example, we might suppose simply that mental properties have been defined just as those through which mental expressions refer, and that we already have some logically *independent* means of determining which expressions are mental referring expressions. Thus, we are now able to accept that mental properties can at least be picked out determinately as those properties through which mental referring expressions refer.

Metaphysically, of course, the demarcation just adduced is still singularly uninteresting. In order ,to inject some metaphysical significance into the distinction, then, we must assume that some further property or characteristic X can be cited which only mental properties have. If a suitable property can be found, we will then be in a position to infer that there is a metaphysically significant distinction to be drawn between mental and physical properties, and that a version of [QD] is true. But if there is such a characteristic X, [QD] stands or falls on the credentials of that property ascription alone, and the property dualism thesis is rendered completely redundant. For if, in general, the uAPC condition can obtain for expressions which refer via *physical* properties, the obtaining of the uAPC condition in a case involving a mental referring expression tells us nothing about the metaphysical status of the property through which it refers. That property will be non-physical just if, and in virtue of the fact that, it bears characteristic X.

This observation leaves our current attempt to support [QD] substantially back at the point of departure in chapter III. Thus, in that chapter [p 96], we saw that Smith can be in an epistemic situation with regard to a phenomenal property Q such that:

1. X is the property of Q, such that Q is identical with P, and in epistemic situation Eq it is possible to pick out Q determinately, but *not* to determine that Q is a PPD property.

Here, an epistemic situation is taken to be a situation in which only a limited body of information is available. Hence, if a fact can be logically inferred from the information in Eq only if supplemented with further information, we can say that it is not possible *in Eq* logically [i.e., a priori] to infer that fact. Thus, in such a case, we would say that the relevant fact can be established only a posteriori, by supplementing the information in Eq with further information. In the present discussion we are assuming the uAPC condition in premise 2 that:

(a) There is an epistemic situation Eq in which a mental referring expression such as "Smith's pain at t" cannot be known *a priori* to co-refer with any PPD referring expression whatever.

And from what we have just said about Eq it then follows that:

(b) There is an epistemic situation Eq in which a mental referring expression such as "Smith's pain at t" cannot be known *at all* to co-refer with any PPD referring expression whatever. [Knowing the fact a posteriori would amount to being in some other epistemic situation].

If the referent of "Smith's pain at t" is p, and P [which is identical with p] is taken to be the referent of any PPD referring expression whatever, it is then apparent that p in (b) has the property X' such that:

> X' is the property of p, such that p is identical with P, and there is an epistemic situation Eq in which it is not possible to determine that p is a PPD property.

If we then assume that the epistemic situation in question is such that in Eq it is known that the referent of "Smith's pain at t" is p, we get:

> X' is the property of p, such that p is identical with P, and there is an epistemic situation Eq in which it is possible to identify p even though it is not possible to determine that p is a PPD property.

And this is clearly just a restatement of the characteristic X described in chapter III. We have as yet found no compelling reason for inferring from the occurrence of X that any non-physical properties are implicated, and while the property dualism argument has been shown to depend implicitly on the occurrence of X as an assumption, we can infer from our discussion in this chapter that it has been shown to offer no additional support for the dualist's thesis.

Conclusion.

We have seen that White's argument fails to establish that epistemic access to a brain state R introspectively must proceed via a non-physico-dispositional property, and hence that some further characteristic X of mental properties is needed to achieve that goal. If, indeed there is such a property, we have yet to find it. What is now certain, however, is that neither Kripke nor White has been able to produce a suitable candidate.

Setting the property dualism argument aside completely, then, the picture we arrive at is by now a familiar one. The question of whether there really are properties which set the mental apart in some metaphysically significant respect from the physical has simply not been addressed. The proposal that mental properties alone are directly introspectible, for example, or that they alone are in some sense epistemologically private, still remains completely unsubstantiated. It is suggestions such as these that White presumably regards as the absurdities which lead him to reject the possibility of mental properties altogether. The property dualism argument itself, however, has nothing to say in this respect.

White's further attempt to discredit the a posteriori mental/physical identity thesis can now be seen as just a more general statement of the epistemic situation in terms of which Kripke's modal intuition was explained. Thus, we might assume with

Kripke that the referent of M, "Smith's pain at t", can be determinately identified without the need for mediating properties of any kind. In that case, however, White is merely subscribing to a more general case of Kripke's epistemic situation in which there might be such mediating properties. The fact that M cannot be known a priori to corefer with any physical expression A must be employed to establish that properties of distinct types are implicated; the only difference being that while for Kripke the mental property is the referent of M itself, White allows that it might not be.

Irrespective of whether this comparison is justified, however, we can now see that White's argument goes through only on an equivocation over the meaning of "logically distinct properties". For even if it can be argued that the a posteriori condition can only arise when two distinct *tokens* of a property provide the routes to the referent, it remains possible that those distinct tokens are each of physical *types*. And since White needs to show that property dualism involves properties of interestingly different types, one physical and the other mental, it follows that he is not entitled to the conclusion he requires.

Nevertheless, he might still be tempted to argue that the a posteriori identity thesis in question is more demanding than we have so far acknowledged. Thus, he might point out that in accordance with that thesis it is only if two expressions refer to a single *token* of a physical referent R that they are to be regarded as coreferential. After all, if a mental expression refers to a physical state at all, there must be some physical expression which refers to that very same token state. But we can readily concede this point, since it renders White's position even weaker. For whether two purely physical properties provide numerically distinct epistemic routes to a single *token* of a physical state must a fortiori be a matter for a posteriori investigation. Thus, for two referring expressions to refer to a single token referent they must, logically, refer to tokens of a single type. And since the latter state of affairs is knowable only a posteriori, it follows that so is the former. Hence, it is logically possible that two expressions should refer via purely physical properties to a single *token* of a brain state R, and yet that this fact should be knowable only a posteriori.

Chapter VIII

UNRESOLVED PROBLEMS

We have found that each of the proposed counterarguments to reductive physicalism depends for its force on quite distinct claims about qualia. Hence, our initial appraisal of each argument took the form of an analysis of the claims being made. Only then was it possible to see whether there is even a problem. We take it that there is at least a *prima facie* challenge to be met if our generic brand of reductive physicalism appears to be in trouble, and in the introduction we outlined the principal expectations we might reasonably have of any reductive programme. Reductive physicalism was cast minimally in commonsense terms as the claim that all occurrent states, properties and events are both epistemically and cognitively available from within the scientific framework of a third person perspective. Taking current science as our initial arbiter of the physicalist's ontic commitments, we needed to find out whether there are any occurrent qualia which appear to be excluded from that ontology. If there are, we might try to find out whether it is still plausible to regard them as physical properties.

The Dualist's Strategies.

The various strategies adopted with respect to the ontic commitments of current science should now be familiar, and for ease of reference we have summarised those strategies in the introduction. Here, we need only provide a brief summary of the relevant findings. One crucial finding was that, in view of the elusiveness of the distinction between reductivism and eliminativism, any intelligible refutation of reductive physicalism would have to be framed in terms of some intelligible characteristic X which qualia have but no physical phenomena with which they might be plausibly identified have. It was reassuring to find, then, that each of the strategies proposed by QD does cite at least some intelligible characteristics of qualia which might be problematic for the physicalist. So in accordance with this strategy the dualist must now find some way of evaluating his intuition that the intelligible characteristic X which he ascribes to qualia and regards as being problematic do

belong to occurrent items but not to the relevant physical items. He can then refer topic-neutrally to the items which have X as "qualia".

The next finding I want to focus on is the observation that none of the counterarguments provides *conclusive* evidence that there are qualia in addition even to the agreed members of [S]. Thus, in the inverted spectrum argument, we found no compelling reason to suppose that there are any qualia which can vary against an entirely fixed physico-dispositional backdrop. Similarly, the knowledge argument left us still wondering whether there are any facts about sensory experience which are not just paradigmatically physical facts. The modal argument was no more successful in this respect; for whether a completely topic-neutral account of what we identify in introspection as a pain, for example, is complete depends on what is actually discerned in introspection. In all three cases, then, the initial complaint might be that the counterargument to reductive physicalism simply begs the question which it sets out to answer, but we shall see now that this would not be entirely accurate. In order to explain what I mean by this I shall take the knowledge argument in particular as my paradigm. What I have to say about it might be applied in parallel fashion to the other counterarguments.

<u>The Knowledge Argument</u>.

In the case of qualia, the strategy here was to show that a complete *knowledge* of the physical facts does not include or entail a knowledge of qualia. In order to avoid simply begging the question, then, the argument can only carry any force if we can cite a plausible criterion for the physical and then show that there are occurrent qualia which fail to satisfy that criterion. To see how this works, we should remind ourselves firstly of the basic argument, which might be condensed into the following format [adapted for present purposes from Robinson's version, 1993, p 163].

1. All and only physical facts [FP] are capable of expression within the vocabulary of physical science. Call this capability EP. Hence, for any x, x is an FP iff it has EP.

2. Smith knows every x which has EP.

Therefore,

3. Smith knows every FP.

But,

4. Smith does not know fact Q.

Therefore,

5. Q is not an FP.

And assuming there to *be* a fact Q, it follows that there is a non-physical fact. In this form the argument is clearly valid, so in order to assess its soundness we need to consider whether the premises are true. We have already seen that in the particular epistemic situation envisaged for Smith premise 2, or its equivalent, has come under scrutiny for a number of reasons. Thus, it seems by no means clear to some commentators that Smith *would* have all the physical facts, or know all there is to know. For present purposes, however, we shall simply assume that he does; that the knowledge argument cannot be charged with question-begging at premise 2.

Since 1 is being taken axiomatically as an indication of what we *mean* by physical facts, furthermore, it is to be assumed uncritically that 1 is true. For the time being we shall ignore the question of what fact Q is, and whether it is true of any occurrent qualia, and assume just that *there is a fact Q*. And this leaves 4 as the only remaining premise.

Justification for Premise 4.

Premise 4 entails that fact Q does not have EP; that it cannot be expressed within the vocabulary of physical science. If he were to assume that 4 is true without further substantiation, however, I would take it that the dualist is just begging the question; it is unsatisfactory simply to assume as a premise that Q cannot be

expressed in the vocabulary of physical science. Hence, however obvious the underpinning for this assumption might seem to the dualist, it must *have* an underpinning of some sort. We can assume, then, that the underpinning must take the following general form.

It can be inferred that Q does not have EP from the fact that:

4' There is some characteristic P, such that Q has P, but no facts which have EP have P.

Substantiation for 4' can be sought in a number of directions, depending on what P might be. Thus, it could be claimed that P is the characteristic of being either:

P1. Not contained in the set [FP] of currently known scientific facts.

P2. Not capable of being contained is the set [FP]' of *future* known scientific facts.

or some characteristic which can be known as true of all possible physical facts but specified without explicit reference to science, such as:

P3. An essentially subjective fact.

It seems clear that the attribution of P1 to Q might at least be testable by anyone who has a full understanding of current science, and who knows what to count as a currently known scientific fact and what not. But while this would enable the dualist to test his thesis about Q in a more or less determinate fashion, it seems that he does not yet have the requisite knowledge or understanding. So it remains possible that Q might yet turn out to be one of the scientific facts already known, but not yet known to *be* that fact. If the knowledge argument is designed to establish that *Smith is in Pain at time t* is not just the fact that *Smith's C-Fibres are being stimulated at time t*, for example, we should expect it produce a reason for supposing that it is not. Again, it is surely true that whatever Q happens to be - call it the fact of *what it is like*, experientially, to see red - current science does not *explicitly* give an account of Q. To suppose otherwise would be, effectively, to already accept

that the topic-neutral account of facts discerned in introspection is *known* to be exhaustive. But while this assumption is unwarranted, future science might nevertheless find it to be true. Indeed, if we were prepared to reject this possibility without further argument, I would again take it that the crucial question was being begged.

But this implies that P2 cannot be acceptable as it stands either. For if, *ex hypothesi*, we do not yet know all the scientific facts it is in principle possible to know, we would require some independent argument to substantiate the claim that Q can never turn out to be one of those facts. So even if Q is thought to be neither P1 nor P2, we still need some plausible support for this thought. And this leaves the dualist with option P3; some characteristic P which all physical facts *must* but which Q does not. I take it that Robinson will accede to my verdict here, since he is at pains to point out that our limited current scientific knowledge is not relevant to the argument [1993, pp 162 - 3]. What *is* relevant, or so we both seem to think, is whether Q is a fact about "the subjective dimension" [op.cit., p 163] or, in my account, an essentially subjective fact. Indeed, it seems from our preliminary outline of what a physical or scientific fact would have to be that in the absence of further specifications for *the physical* this is the appropriate issue. For we characterised physicalism in commonsense terms just as dealing exclusively with facts knowable in the third person perspective; facts belonging to the *objective* dimension. So we can take it that even if the project is to show just that *current* science is not ontically committed to qualia, even under any other name, the strategy will be to cite some characteristic which distinguishes Q from all physical facts and thence, a fortiori, all currently known physical facts.

<u>The Subjective/Objective Distinction</u>.

The proponent of the knowledge argument appears to be committed to the strategy of showing that there are facts about qualia which are only subjectively knowable. And this is just because we are entitled at this stage to assume that any facts which are *objectively* knowable, in the third person, might be physical facts. So we might rephrase the claim about qualia facts as the proposition that Q is

not objectively knowable. The first question, then, must be "what would it amount to for a fact to be objectively knowable?".

It can be assumed in deference to the dualist that there is a subjective way of knowing some facts, and that to know a fact subjectively does not entail knowing that fact objectively. If this were not true, the dualist's claim that Q is knowable *only* subjectively would be vacuous, since there would be no possibility of knowing a fact subjectively but not objectively. Conversely, and by the same token, we must also assume that knowing a fact objectively does not entail knowing that fact subjectively. If this were not true, knowing a physical fact objectively would entail knowing that fact subjectively, and it would not be possible to say that a physical fact can be known without knowing it subjectively. For ease of reference, then, we might refer to the two logically distinct ways of knowing facts as being by:

K1. Acquaintance - roughly, by direct conscious experience,

K2. Theoretical understanding - by any means *other than acquaintance* available in principle to the physicalist.

and from what we have just said, it will be assumed that:

K1'. K1 does not entail K2.

K2'. K2 does not entail K1.

Replacing subjective and objective with the labels K1 and K2 respectively, we can then say that the distinguishing feature of Q is supposed to be that it can be known *only* by K1. But this presents the dualist with a rather serious problem. Again, we can take the knowledge argument as our paradigm.

In brief, the argument might now look like this. Since (i) each particular physical fact can be known by K2, and yet (ii) Q can be known only by K1, (iii) Q is not a physical fact. Consider, then, how the knowledge argument in this form might be substantiated.

As before, we might take premise (i) as being true axiomatically. In that case, all that would remain would be to establish that premise

(ii) is also true; that Q can be known only by recourse to acquaintance. So how might (ii) be substantiated? The problem is that, as we have already observed, the dualist is not entitled to *presuppose* that Q is not a fact already known by current science, or a fact which might yet be known by science. In order to rule out either possibility, then, he is obliged to cite some characteristic P which Q has but no physical fact *can* have. Clearly, he cannot simply cite P as being the characteristic of being knowable only by recourse to K1, since that would amount to begging the question as to whether (ii) is true. But the problem is more fundamental. For not only is it not possible at present to presuppose that Q is not just a physical fact already known, but it seems that we do not even have a reliable *criterion* by which to pick out instances of knowledge by K1.

The dualist has a ready answer to this problem, however. He will insist that it is just obvious that what I am referring to by my "being in pain" is a fact which can only be acquired by acquaintance, because it must be acquired in *introspection*. Hence, there must be something wrong with my argument. For then it is just obvious that it is possible to tell when knowledge is essentially acquired by K1 rather than K2. This objection would miss the point, however. For the point is that what we find *just obvious* might turn out to be false; the knowledge we take to be gained only in introspection might be just physical knowledge. To suppose that this is not so would be to presuppose that the topic-neutral analysis of what we refer to as our "introspection of pain" is false. So if K1 is simply *defined* as the way in which non-physical facts are known, we have not yet established what that way is.

It is difficult to see how the definition of K1 might be tightened up to ensure that all knowledge acquired by K1, or "introspection", must be non-physical. For even if we say that K1 must be by immediate introspective access to "*raw feels*" in particular, we have yet to establish that "a raw feel" is not just a topic-neutral reference to a physical phenomenon. To suppose that we have established that would amount to begging the question which the knowledge argument set out to answer. Thus, the argument was of the form: since (i) each particular physical fact can be known by K2, and yet (ii) Q can be known only by K1, (iii) Q is not a physical fact. But now K1 has been *defined* as knowledge of raw feels by

introspection; so now premise (ii) becomes (ii)' Q can be known only by K1, because that knowledge can be gained only by introspective knowledge of raw feels. If (ii)' is true, then, from (i) and (ii)' we can draw the required conclusion. But it would beg the question at issue to suppose that the knowledge of *raw feels* acquired in introspection cannot be acquired in any other way.

If an appropriate form of knowledge K1 is to be found, then, it must satisfy the following criterion. It must be possible to know that any knowledge gained by K1 will be non-physical knowledge logically prior to knowing that anything known *is* non-physical. Otherwise, the knowledge argument will beg the question it sets out to answer. I have not claimed to have shown that no such criterion can be met; only that it remains to be seen *how* it might be met. This is a requirement which must be met if the knowledge argument is to be taken seriously.

The only way in which the dualist might circumvent the above assignment would be by attending instead to a mode of access to *physical* knowledge which cannot provide knowledge of raw feels. Thus, although he might concede that he is unable to meet the criterion for K1, he might still claim that he has an effective criterion for K2. He might suggest, for example, that although physical facts *can* be known by introspection, they need not be. The very notion of the physical entails that all physical facts *can* be known in the third person perspective, and whatever introspection might be it certainly does not provide that perspective. So if K2 is taken to be something like "knowledge acquired in the third person perspective", we might yet find a convincing version of the knowledge argument. It will then run as follows. (i) all the physical facts can be known by K2, but (ii) Q cannot be known by K2, because *that* knowledge can only be gained by introspection. Therefore, (iii) Q is not physical. This seems promising, because the argument is again valid, given what we have just specified about introspection, and K2 at least plausible. But the question remains as to whether the premises are true. And we can perhaps see that the premises taken together must be supported by a further assumption; that all physical facts, but not Q, can be known without introspection. But this is just a version the premises in the argument already given [p 255], taking K1 to be *by introspection* and K2 to be any *other* way. So if it is true, the conclusion will follow

anyway. Whether we try to clarify K1 or K2, then, the problem is the same; to specify a way of knowing which is essential for Q but not for physical facts. So we need to specify both an appropriate Q and an appropriate way of knowing Q.

The topic-neutral approach is intended to show why this need cannot be met. Thus, if it can establish that "Q" is just an unwitting reference to a physical fact which we could know in K2, we will be forced to concede that there is no *appropriate* Q after all. For example, if it can be shown that "Pain" is an unwitting reference to C-fibre stimulation, we would then be unable to claim that pain is non-physical, or that *my being in pain* is not a physical fact. For then my being in pain is nothing more than having my C-fibres stimulated.

The Topic-Neutral Strategy.

The concept of epistemic topic-neutrality can be explained with the help of a model, and we can initially suppose that it is a *purely physical model*. Firstly, we can assume that the identity relation in question amounts to the candidates A and B being one and the same state or property. Thus, it cannot be just that A an B are of a single *type*, since that would entail only that they share a common property or attribute. Whether A and B are properties [universals], tokens of a property, or individual objects or items, then, they must be one and the same. So for the sake of the argument we shall assume that the candidates in question are *individual items*. In the mental/physical corollary this would amount, say, to A being a particular pain, and B being a particular C-fibre stimulation. In our physical model, then, suppose that Smith is standing at the mouth of an estuary, E. A little way upstream the estuary divides into two rivers, and further up each of those rivers divides into two streams. There will be be four streams in all, then, which we shall refer to as S1 ... S4 respectively. Four exactly similar objects, O1 ... O4, are dropped simultaneously into each of the streams, and allowed to float down to the estuary, where Smith is then able to pick out O1. But he clearly does not *know* that it is O1. He only knows that it is one of the objects E1 ... E4, and somewhat fortuitously labels it as E1. So we can say that although E1 is identical with O1, Smith can only identify it as E1; the

particular object he picked out of the estuary. Epistemically, then, his determination that it is E1 is *topic-neutral* with respect to O1. O1. He has sufficient information to determine that it is E1, but not to establish that it is O1. But since there is no further information to be had at the estuary, he would have needed a bird's eye view of the overall situation in order to determine that E1 *is* O1. So we might refer to the [epistemic] situation at the mouth of the estuary as the first person perspective, and that in the bird's eye view as the third person perspective. It then follows that in the first person perspective Smith can determine that the object is E1, but not that it is O1. Since this is a purely physical model, however, we have no reason to suppose that this latter fact cannot be ascertained in the third person.

Still assuming that only physical states are involved, the model can now be applied to Smith's token state CFS. Thus, in the first person perspective he can pick out CFS as an individual state, but does not *know* that it is CFS. So he refers to it instead as Pain, the token state he can pick out. Since the entire set-up is couched in physical terms, however, it must be assumed that from the third person perspective he could determine, at least in principle, that Pain *is* CFS. So, just as before, we can say that Smith's epistemic situation is topic-specific with regard to Pain, but topic-neutral with regard to CFS. In order to pick out Pain topic-specifically, however, he cannot pick it out in virtue of its having any properties; for that would entail a topic-*neutral* identification of Pain; as whatever state has those properties.[47] For suppose to the contrary that he *did* pick out Pain just as whatever state has property P. It would then follow that P is epistemically topic-neutral with respect to both Pain and CFS, even if Pain is topic-

---

47. It is *possible* to assume that even Pain is not known topic-specifically, of course, but this would amount just to the identity of Pain and CFS being knowable only a posteriori. This position was explored in chapter VII. The point of the present discussion is to comply with the dualist's intuitions as far as possible. And one of his intuitions is that it is possible to know Pain topic-specifically in introspection. So the question must be whether the topic-neutral account of perception can accommodate that intuition.

specific with respect to CFS. And in that case we would have failed to explain what it amounts to for *Pain* to be topic-neutral with respect to CFS. In order to understand the latter relation, then, we must understand Pain topic-specifically. And since, ex hypothesi, Pain and CFS are one, this entails being able to pick out CFS determinately in introspection, but without knowing that it *is* CFS. That topic-neutral recognition of CFS as Pain must have been achieved by way of topic-specific recognition of Pain.

So our topic-neutral account of picking out CFS in the first person presupposes that CFS is picked out topic-*specifically*, but not *as* CFS. We can see this more clearly by referring back to the estuary. Suppose that in addition to referring to the object Smith chose as E1 and O1, we also refer to it as R1; the object which travelled down the first river. In that case, Smith knows *both* O1 and R1 only topic-neutrally, as E1. This is epistemically parallel to Smith knowing both Pain and CFS topic-neutrally as the bearer[s] of P. Given this state of affairs, then, it matters little whether he knows R1 as O1. What we should say, rather, is just that even if he knows *that* R1 is O1, he is epistemically only able to pick out R1/O1 topic neutrally as E1. So we can explain the topic-neutral epistemic relation in terms of Smith being able to pick out the object as E1 topic-specifically; but not as R1 or O1. Similarly, then, if the topic-neutral epistemic relation obtains between Pain and CFS it does so in virtue of Smith knowing Pain topic-specifically as Pain, but only topic-neutrally as CFS.

It seems clear that for the reductive physicalist the deficiency in Smith's knowledge *can* only amount to his not knowing that the particular token he knows in the third person as CFS is the particular token he knows in the first person as Pain. For we are entitled to assume that in the third person he has access, at least in principle, to all the facts about CFS, and therefore about Pain, other than the relational fact just mentioned. It is just that while he is observing Pain in the *first* person, he might not have access to all of those facts. And this is a mixed blessing for reductive physicalism. It is good news in that such an epistemic situation is clearly compatible with the physicalist's position. Thus, even our metal friend can be in an epistemic situation in which he is able to know all the facts about a single physical state from two distinct perspectives, and yet not know that it is *one*. As we suggested in

chapter VII, it would be question-begging to assume that he could not. On the other hand, however, this leaves the physicalist's position at its most vulnerable. For if there *are* any facts which *can* even in principle be known topic-specifically only in the first person perspective, we would be able to infer that his position is false.

In the light of our earlier discussion, we can now see that the topic-neutral strategy is both eliminative and reductive. It claims that if the fact to which we refer is non-physical it is not a fact, while if is physical it is a fact. So there are two possible ways of dealing with our reference to qualia. The first is to say that when we think we are experiencing a quale our belief is just false, and the other that what we believe in is just a physical fact. But, as we noted earlier, it is not always clear *which* of these verdicts is being passed on our qualia. Taken in the eliminative sense the denial that we experience qualia can seem preposterous. Thus, Peter Smith finds himself in the characteristically Wittgensteinian position [under one interpretation] of having to deny outright that there *are* any qualia.

> ... perhaps all that happens is that we can just 'repeat an expression' - i.e. say straight off, without relying on observational evidence at all, whether we are in pain or not [1986, p 206]

If what he means is anything *like* what he says we can infer that for him there are no sensations of pain. Smith is a functionalist, but a type-identity theorist can find himself in an equally puzzling position. Hill, for example, is adamant that when someone seems to be aware of something in an hallucinatory state:

> ...these appearances are misleading. .... To be aware of a sensation it is necessary to be aware *that* some proposition is true. ..... But prior to the moment of forming the belief, he is not aware of anything - however much it may seem to him otherwise. For, prior to that moment, he has not activated any concepts that stand for sensations. [p 195]

Even if this *is* a version of type-identity, it effectively eliminates the sensation types and talks instead about sensation *concepts*. And if it is necessary to form the concept in order to "have the sensation" we might wonder how he was able even to form

the concept; for if he was not aware of *anything*, he presumably had no information [see chapter V]. Smart also seems to recognise that the question of whether there are any subjective experiences *per se* deserves attention. He says in what *might* be seen as an eliminative turn that:

> ... in so far as a sensation statement is a report of something, that something is in fact a brain process. Sensations are nothing over and above brain processes. [Rosenthal, p 170]

But in the same place he insists that there is a strict *identity* between sensations and brain states [p 171]. So if our reference to sensations is a reference, albeit unwitting, to brain states, there *are sensations*. It is no longer just that there are no sensations *over and above* brain states; they just are brain states. Smart seems unclear as to which of these positions to adopt. Indeed, if his position is that sensations are brain states, we might even suppose that he is in agreement with Kripke. For Kripke's view is that pain is picked out epistemically by its "immediate phenomenological quality" [p 152]. So it could be suggested that Smart's *epistemic* position on pain *allows* that C-fibre stimulation might be picked out immediately, although not *as* C-fibre stimulation, in introspection.

Now the positions just mentioned are all couched in quite different philosophical views on the nature of perception, but this serves to illustrate the present point; that whatever physicalistic position is adopted its treatment of topic-neutrality is a central theme. For unless the account can explain which supposed facts about qualia are physical facts, and which are not facts at all, it will be impossible to apply it meaningfully to the dualist's particular claim.

Just how difficult it is to clarify this issue can be seen from the following, Thus, instead of wondering which facts are held to be physical facts and which not facts at all, we might suppose that it would at least make sense to come down firmly on the side of the eliminativist. For his thesis is that there are *no facts at all* which are knowable only in introspection. But we can see straight away that this fails to distinguish his position from any form of identity thesis. In all cases it is held that all the facts are physical facts, knowable in the third person perspective. So the

problem now confronting the topic-neutral strategy is precisely the problem we explored in chapter III; of how to make sense of a distinction between eliminative and reductive physicalism. Since we were unable to find a plausible distinction there, there is no reason to suppose that we can do so here. All we can say is that sensory perception is topic-neutral in that every fact known in introspection is, but might not be known as, just a physical fact; *knowable* in the third person.

Epistemic Topic-Neutrality.

What the knowledge argument presupposes is that there are experiences, or experiential qualities, which can be known only by direct introspection. For only if we assume that there are phenomena which satisfy this description can we determine that the physicalist's reductive account is incomplete, even in principle. But we can now see that, in turn, the inverted spectrum argument presupposes the *conclusion* of the knowledge argument. For if there were no occurrent properties in addition to the complete set of physico-dispositional properties, it *could* not be true to say that there are any properties or states which can vary against a completely fixed physico-dispositional backdrop. Varying any properties or states would *amount* to varying the backdrop. The modal argument we considered is no-better placed. In the absence of any further information about qualia, Kripke's illusion of contingent identity was explained away in purely epistemic terms, and White's epistemic observation turned out to offer nothing more persuasive either. In both arguments, we found that it would be necessary to presuppose that the topic-neutral analysis of qualia and experience is false in order to derive anything of interest. And again, we might assume that the topic-neutral analysis can only be false if there are properties or states which are epistemically available only by direct introspection. So the last two arguments also presuppose the conclusion of the knowledge argument.

The crucial point here is that *if*, or *insofar as*, it is like anything at all to experience a pain, then for the reductive physicalist it must be possible to find out what it is like by making third person observations; without *having* the pain. According to the topic-neutral account, every fact knowable in introspection

is also knowable in the third person. But it is extremely difficult to imagine how this could be possible for pains. Indeed, it is difficult to accept that it could be like *anything in particular* to observe another person's states of pain. For there are various physical modes of access to those pains, and each mode will be quite unlike the others. And this appears to be true for other physical states. What, for example, do reading a thermometer through the window and standing out in the cold have in common? In each case we acquire the information that it is cold, but there are no other obvious similarities. And the information we thus acquire in each case leads us to the same knowledge about the physical state of the air. So in each case we can be said to learn all the relevant physical facts, although there is nothing that it is like in *both* cases to do so. So if the physicalist tries to accommodate the dualist's intuition by insisting that the particular experiential character of pains can be known through third person observation it seems that his position will be incoherent. For if, in general, it is not like anything *in particular* to know a physical state in the third person, how can it be so for pains?

It is a feature of the physical world that cognitive access to [i.e., the ability to know and understand] all the physical facts can be gained in a number of quite diverse epistemic ways. In view of this, there seems to be no other sense in which we can have third person epistemic access to physical states or properties. Epistemic access affords a route to a complete knowledge and understanding of the physical facts, and is itself "complete" just if it fulfils that role. And this should come as no surprise, when we consider the conceptual content of theoretical physics. What, for example, would a complete knowledge and understanding of a magnetic field amount to? The radical underdetermination of the physical facts by the sensory evidence is now taken for granted; so we should not *expect* it to be like anything in a particular to have that knowledge and understanding. So *epistemic* access to the physical might be construed just as sensory access to a complete cognitive grasp of the physical facts, for there is no other sense in which we might expect to *know* those facts. And if this is true of physical states of an uncontroversial sort, we should surely expect it to be true of pain. Hence, if the reductive physicalist's account of pain is complete, his only coherent position must be that there is nothing in particular that it is like to have a pain. If it were, we would

not expect to have epistemic access to what it is like in the third person. And this entails for him that our introspective ability to discriminate pains is unaccompanied by any particular quality; that discriminating pains amount *just* to the indeterminate discrimination and understanding of physical states.

The situation *in fact* seems quite different, however, and it is doubtful whether this implication of reductive physicalism can be true. For we began by ordaining that what we mean by a pain, or token E[Pain], is a particular state which we can discriminate determinately in introspection; a state which it is invariably like something in particular to be in. Irrespective of what it *is* like, then, a state of pain is something which can be determinately discriminated in the first person, without recourse to any theoretical support. So if our discrimination of magnetic fields, or of the atomic weight of a Caesium atom, entails our cognitive grasp and assumption of the appropriate theory, it would seem that pain discrimination is quite unlike the discrimination of any such physical facts. Pain just *is* of a type which can be determinately discriminated in introspection. And if a state of this type *were* a third person observable, it would leave pain out on a limb with respect to most physical states.

This clearly leaves the reductive physicalist with a serious conceptual problem. For the discrimination of pains in the third person is in fact *very much* like the discrimination of magnetic fields, or of any other physical phenomena which cannot be picked out determinately in the way that first person pain sensations can be picked out. Thus, we have yet to formulate a plausible answer to the question of how "this soggy grey matter" of the brain [McGinn, 1989, p 349] can possibly produce the sensation of red or of pain. More specifically, we cannot understand how *any* introspectible types can be physical types, in view of the fact that introspectible types can be determinately discriminated to a degree which physical states in general cannot. So if we really can discriminate sensation types to a degree that is not possible for physical types in the third person, it is difficult to see how the physicalist might provide a complete third person account of our pains. For there seems to be a clear asymmetry between the degrees of determinacy available in the first person and third person epistemic perspectives. Furthermore, even if we have the *false belief* that we experience those particular

and qualitative sensations which seem so compelling, it is still difficult to see how the epistemic *belief*-types of first person experience can be translated into the third person types of physical science.

It seems, however, that there might be a plausible rejoinder to this line of argument. Thus, the reductive physicalist could say that the reason why we can discriminate sensation types so crisply in introspection is that the types we discrimate thus are the result of our particular physiological make-up. So the reason why we find pains so easy to discriminate is that in so doing we are simply responding to physiological states of a particular type; stimulation of our pain receptors. We do not have magnetic-field or ionised plasma receptors, so we cannot discriminate those phenomena very crisply at all. In fact, quite a lot of instrumentation and theory is required to do so. But if we had a magnetic field sensor built into us our discrimination of magnetic fields would be as crisp as it is for pains. And, of course, it is not that our pains are caused by any *particular* type of external physical stimulus; *anything* which stimulates the receptors will be discriminated, topic-neutrally, as being painful.

So there is at least a prima facie plausibility about the suggestion that what we are able to discern in introspection as a discreet type might not be discernable as such in the third person. For a discreet type discernable in introspection might be merely a function of the various kinds of external stimulus our receptors respond to. Another example might help to clarify this. We know that we can discriminate red stimulation quite cleanly; that our red cones respond [more or less] only to red light. Similarly for green and blue. When irradiated with yellow light, a similar discriminatory ability is apparent. But what happens when we are irradiated with red and green light simultaneously? Again we discriminate it as yellow; or, at least, we discriminate our experience as being of the sort typically produced by yellow. And the explanation for this phenomenon is perfectly simple. It is that the visual set-up in our brain is organised in such a way as to recognise when the two cones, red and green, are being stimulated at once. So although yellow light has nothing in common with red or green light, as far as wavelengths are concerned, the same epistemic discrimination is made in response to the two quite different sorts of stimulus. In each case we will

discriminate 'yellow'. And from such simple examples it seems perfectly clear that the types available to us epistemically are quite unlike the types of stimulus that produce them. So, for the reductive physicalist, it will seem perfectly natural that pain, or yellowness, is so much easier to discriminate than a magnetic field. And it will also seem perfectly natural that the state of being in pain, or of discriminating yellow, will be difficult to discriminate from a third person point of view. For from that vantage point the observer is not discriminating yellow; rather he is observing and analysing the *complex neural set-up* involved in discriminating yellow. That neural set up might be quite as difficult to discriminate as a magnetic field after all.

Although this explanation of sensory discrimination seems to be casting a favourable light on reductive physicalism, however, for the dualist it will be just obviously false. For even though it provides a plausible account of how a *physical* being might have discriminatory abilities just like ours, it simply misses something out. And the dualist might just sit tight and insist that seeing yellow *is* more than just having the appropriate discriminatory ability. For whatever seeing yellow amounts to, it certainly seems to involve a qualitative experience of some sort. And, as we have seen, it is unhelpful to explain that what we believe to be a qualitative experience of that sort is not really there, or that our belief that there is such a qualitative experience is just a false belief. For the dualist will either argue that belief *per se* is incapable of a purely physico-dispositional analysis, or that having the intelligible concept which is invoked in that belief is incapable of physico-dispositional analysis. At the very least, he might insist that it seems incomprehensible that a belief is all there is to it. For it seems utterly implausible to say that seeming to see colours in the way we do seem to see colours amounts to nothing more than a neural state, either neurally or dispositionally characterised.

<u>The Topic-Neutrality of Physical Knowledge</u>.

Even if the dualist is right to reject that suggestion, however, there is a further and more radical proposal which the physicalist might make. Thus, even if there is more to sensory perception than

the physicalist's reduction allows, he might try to accommodate this fact by casting doubt on the extent of our possible knowledge in the *third* person perspective. So when we say that we can *know* CFS in that perspective, for example, what is it that we can really know? It might be plausibly maintained that since such knowledge is mediated by our senses and by scientific method, and couched in framework of physical theory, there is a sense in which we cannot know CFS *per se* at all. Russell's *neutral monism* can be interpreted as proposing this sort of approach.[48] The particular point of interest in the present context is just this. If we can make out a plausible case for claiming that our knowledge of the *physical* state CFS in the third person must, even in principle, be incomplete, then it might no longer be a problem for the physicalist if our knowledge of pain in the first person contains *additional* information. For then that additional information might still be information about CFS *per se*.[49] This proposal is beyond the scope of the present discussion, but we can at least see how it would affect the physicalist's position in principle. For if it were true, there might be no difficulty in explaining why there seems to be experience, or sensations, which bear no obvious relation to physical states at all. The sensations just are what we pick out,

---

48. The position is developed in Russell, 1927. Later attempts to develop along similar lines are in Foster, 1982, where the resultant position is nevertheless quite different, and Lockwood, 1981, 1989 and 1993.

49. One of the most significant challenges to the thesis was spelt out as the *grain problem* by Wilfrid Sellars [1965, pp 430 - 51]. The essence of the challenge is that there appears to be no plausible way of explaining how some of the intrinsic properties of sensations might be physical properties. Seeing a coloured patch, for example, my sensation is an extended *homogeneous* feature of experience, whereas nothing in the brain, as known in the third person, seems to have that characteristic. A more recent version of this position appears in Foster, 1991, pp 126 - 30]. Michael Lockwood suggests a possible way of overcoming the difficulty in 1989, and 1993, pp 271 - 91 within the conceptual framework of quantum mechanics.

topic-neutrally in the third-person, as physical states or properties.

The position just described is not without its problems, however. We have already cited the grain problem as a difficulty for the reductive physicalist, and we can readily see how it might also be a problem for Lockwood. For if states and properties of the brain *per se* really are just qualia or consciousness, discernable topic-specifically in introspection, it would seem that we are forced to concede that what we take to be physical items [atoms, electrons, or whatever] of recognisable physical types are in fact completely different *introspectible* types in different contexts. Thus, it seems clear from science that all electrons, for example, are fundamentally of the *same type*, while according to Lockwood's proposal some would be visual qualia, others pains, and yet others presumably nothing introspectible at all. And it is not at all clear how this disparity might be explained by Lockwood. For it is difficult to imagine how such a wide variety of distinct introspectible types could be analysed in terms of a relatively small number of discreet physical types. At the very least, we can say that insofar as this is a problem for *reductive* physicalism, it is also a problem for the *topic-neutral* account of the physical.

In order to render the identity thesis more plasuible in this respect, then, it seems that it would be necessary to identify the introspectible types with more complex, and therefore multifarious, physical or dispositional configurations or with the behavioural attributes of atomic components. Sensations of pain will be complex atomic configurations of one type and sensations of red another, even though atomic constituents of a single type are involved in each. But while this strategy might accommodate the evident diversity of the sensation types, it seems likely that it will fall foul of the grain problem. Swarms of atoms or electrons have a structural and behavioural complexity which bears no obvious relation to the apparently monadic character of sensations. So at whatever level of macroscopic or microscopic analysis the identity is supposed to obtain, we will have to explain away one problem or the other. And it is not at all clear how this might be achieved. The diversity of experiential types has no evident counterpart at the level of fundamental physical particles. On the other hand,

however, the structural simplicity of a particular sensation simply does not correspond with the complex structure exhibited by a *cloud* of atoms or electrons. So although we cannot say that this problem affords a conclusive objection to the identity thesis, there is clearly considerably more work to be done. Even if a plausible case could be made out for that identity thesis, however, this would still not entail that *physicalism* is true.

Thus, Lockwood acknowledges, rightly I think, that such an account of sensations and sensation types would be *non-reductive*. For if a sensation of pain is only knowable topic-neutrally in the third person, there is nothing to which it might be topic-specifically reduced. Sensations and sensation types are left irreducibly specifiable only within the first person conceptual framework; as the sensations and sensation types which can be discriminated epistemically only in introspection. But then it is not at all clear that his resultant position is not just a brand of *dualism*. For even if physical phenomena and types are discernable only topic-neutrally in the third person, they are so discernable in virtue of the *appearances* they present in that mode of knowing. Knowing a pain in the third person amounts to knowing, topic-neutrally, the sensation which presents as a C-fibre stimulation. So whatever the third person account of those appearances might be, it appears to invoke states or properties which are knowable topic-*specifically* in the third person, and in terms of which those appearance can be specified. Hence, the third person account is not *just* a topic-neutral account of sensations. But this leaves Lockwood having to acknowledge two quite distinct types of phenomenon after all. There are the sensations, knowable topic-specifically only in introspection, and their appearances, knowable topic-specifically in the third person. And this amounts to conceding that although there are phenomena which can be fully known from within the third person perspective, there are others, the sensations, which can not. So unless the appearances are themselves reducible to sensations and their types, which Lockwood appears not to concede, there are still two fundamentally distinct types of phenomenon. And this is just what the dualist claims; that there is an objective or third person realm, *and* a subjective realm.

This leaves the topic-neutral analysis of introspected sensations and sensation types as the only available strategy for the reductive

physicalist. For assuming that there *is* a physical realm, of items and types observable and distinguishable in the third-person perspective of science, a successful reduction of sensations and their types to the observables in that realm will have to leave nothing out. So the crucial question which remains is just whether anything *is* left out by that reductive account, and it seems that there are only two ways in which the physicalist might even hope to establish that there is nothing.

In the first, he must try to show that what we ordinarily take to be experience and its content as discernable in introspection amounts to nothing more than the subject matter of a set of false beliefs or misleading appearances. We have been unable to see how this strategy might work, however, since the analysis of experience as the mere having of beliefs or inclinations to believe tells us nothing about the *subject matter* of those beliefs. And since the subject matter in question is not just a set of propositions about topic-neutral discriminatory abilities, there seems to be no conceivable way in which that subject matter might be rendered intelligible and discernable in the third person perspective.

In the second, then, it will have to be maintained that the subject matter of our sensation beliefs consists entirely of third person observables. But the evident monadic character of our sensations and their types appears to rule out this strategy too. If there really is something in particular that it is like to see red, or to have a pain, we cannot plausibly say that this characteristic of experience is nothing more than a topic-neutral discrimination of physical phenomena and types. For this would be to deny that seeing and discriminating red amounts to more than simply acceding to the proposition that "this physical property is redness"; but it just does amount to more than this. Redness is discriminated as redness in virtue of the particular experiential character it produces. To deny this would be to deny that the concept of redness contains or presupposes the concept of anything which is topic-specifically knowable. But it really does appear to contain the concept of something topic-specifically knowable - the concept of *what it is like* to experience redness.

The difficulty for the reductive physicalist, then, is that we evidently have topic-specific knowledge of sensations for which

there is no parallel in the case of paradigmatically physical phenomena. For the latter are, by and large, only topic-neutrally discriminated as those phenomena and types which our sensory faculties and physical theory enable us to individuate. Magnetic fields, atomic weights and even fundamental physical particles are discriminated in this way. So there is nothing *in particular* that it is like to have those discriminatory abilities. And this must be contrasted with the unavoidable fact that there just is something in particular that it is like to experience sensations. So it surely follows that what we discriminate in experience cannot just be topic-neutrally discriminated physical phenomena. Thus, Jones might discern that Smith is having a red experience in much the same way that he discerns the atomic weight of a sample of Caesium, by picking out properties or appearances of each phenomenon which are determinately recognisable in the third person. But while this might be unproblematic for the physicalist's analysis of atomic weights, it seems irretrievably problematic for the reductive analysis of experiences. For the first person discrimination of Smith's red experience is achieved topic-specifically, and there is at present no conceivable way of *reducing* that experience to physical phenomena knowable topic-neutrally in the third person epistemic and conceptual perspective. At the very least, we can say that if such a reduction were correct, there would be something extraordinary about sensations as physical phenomena. For we would then have to concede that whereas they are like other physical phenomena in respect of being knowable only topic-neutrally in the third person, they are unique in also being topic-specifically knowable in the first. And this appears to leave Lockwood's proposal as the only remaining option - that what is knowable in the third person topic neutrally is reducible to what is knowable in the first person topic specifically. But it is not at all obvious that this would even be a brand of physicalism.

BIBLIOGRAPHY AND ADDITIONAL READING

ARMSTRONG, David. *A Materialist Theory of Mind.* London, Routledge and Kegan Paul, 1968.

'The Nature of Mind'. In C. V. Borst (ed), *The Mind-Brain Identity Theory.* London, Macmillan, 1970.

AUSTIN, J. L. *Sense and Sensibilia.* Oxford. Clarendon Press, 1962.

BACH-Y-RITA, Paul. *Brain Mechanisms in Sensory Substitution,* New York and London Academic Press, 1972.

BEALER, George. 'The Rejection of the Identity Thesis'. In Warner, Chapter 27. pp 355 - 388.

BLAKEMORE and GREENFIELD. (Eds). *Mindwaves.* Blackwell, 1987.

BLOCK, Ned. 'Inverted Earth'. In J. E. Tomberlin, ed., *Philosophical Perspectives,* 4: Action Theory and Philosophy of Mind, 1990. Atascadero, CA: Ridgeview Publishing, pp. 53 - 79.

CAMPBELL, Keith. *Body and Mind.* New York, Doubleday, 1984.

CARRUTHERS, Peter. *Introducing Persons. Theories and Arguments in the Philosophy of the Mind,* Routledge, 1990.

CHURCHLAND, Paul. 'Eliminative Materialism and the Propositional Attitudes'. *Journal of Philosophy* 78, no 2 (1981). Also in 1989.

'Reduction, Qualia and the Direct Introspection of Brain States', *Journal of Philosophy* 82, no. 1 (January 1985). Also in 1989.

Matter and Consciousness (revised ed.) MIT, 1988.

A Neurocomputational Perspective. The Nature of Mind and The structure of Science. MIT, 1989.

DAVIDSON, Donald. *Actions and Events.* Oxford, Clarendon Press, 1980.

DENNETT, Daniel. *Brainstorms.* Harvester press, Hassocks, Sussex. 1979.

'Can Machines Think?', in M Shafto, (ed.) *How We Know.* New York: Harper and Row, pp 121 - 45.

The Intentional Stance, MIT, 1987.

Consciousness Explained. Little, Brown and Company, 1991.

'Living on the Edge'. *Inquiry,* 36, March 1993.

DRETSKE, Fred. 'Mind and Brain'. In Warner, Ch. 10.

FOSTER, John, *The Case for Idealism*, London, Routledge & Kegan Paul, 1982. See also 'The Succinct case for Idealism', in Robinson 1993, Ch. 13.

The Immaterial Self, Routledge, 1991.

GOODMAN, Nelson. *Ways of Worldmaking*. Hassocks, Sussex: Harvester 1978.

HALES, Steven D. 'Certainty and Phenomenal States'. *Canadian Journal of Philosophy*, Vol. 24, No. 1, March 1994.

HARMAN, G. 'The Intrinsic Quality of Experience'. In J.E. Tomberlin, ed., *Philosophical Perspectives*, 4: Action Theory and Philosophy of Mind, 1990. Atascadero, CA: Ridgeview Publishing, pp. 53-79.

HILL, Christopher. *Sensations*, Cambridge U.P., 1991.

HOFSTADTER, D.R. and DENNETT, Daniel. *The Mind's I. Fantansies and reflections on Self and Soul*. New York, Basic Books, 1981.

HUME, David. *An Enquiry Concerning Human Understanding*. Oxford, Clarendon press, 1975. First Published 1748.

JACKSON, Frank. 'Epiphenomenal Qualia', *Philosophical Quarterly*, vol.32. no. 127, 1982.

What Mary Didn't Know', 1986. Reprinted in Rosenthal, pp 392 - 394.

KIM, Jaegwon. 'Psychophysical Laws', in E. Le Pore and B. McLaughlin (eds) *Actions and Events*, Oxford: Basil Blackwell, 1985.

KOLERS, P. A. and von GRUNAU, M. 'Shape and Colour in Apparent Motion'. *Vision research*, 16, 1976, pp. 329 - 335

KRIPKE, Saul. *Naming and Necessity*. Oxford, Basil Blackwell, 1980.

LEVINE, Joseph. 'Materialism and Qualia: The Explanatory Gap'. *Pacific Philosophical Quarterly*, 64, 1983. pp 354 - 61.

LEWIS, David. 'Psychophysical and Theoretical Identifications'. *Australian Journal of Philosophy*, 3, pp 249 - 58.

'Mad Pain and Martian Pain', 1980, reprinted in Rosenthal, pp229 - 35.

LOAR, Brian. "Phenomenal Properties", Philosophical perspectives, 4: Action Theory and Philosophy of Mind. Atescadero, CA: Ridgeview, 1990. pp 81 - 108.

LOCKE, John. *An Essay Concerning Human Understanding*. A. Campbell Fraser (ed.) New York: Dover, 1959. First published in 1690.

LOCKWOOD, Michael. 'What Was Russell's Neutral Monism?', in Peter A. French, Theodore E. Vehling, Jr. and Howard K. Wettstein (eds.), *Midwest studies in Philosophy Volume VI : The Foundations of Analytical Philosophy*, (Minneapolis: University of Minnesota Press), 1981.

Mind, Brain and the Quantum. Blackwell. 1989.

'The Grain Problem', in Robinson, 1993, Ch. 12, pp 271 - 91.

LYCAN, William. *Consciousness*, MIT Press, 1987.

MADELL, Geoffrey. *Mind and Materialism.* Edinburgh U.P., 1988.

MacDONALD, Cynthia. *Mind-Body Identity Theories*, Routledge, 1989.

McCONNELL, Jeff. 'In Defense of The Knowledge Argument', *Philosophical Topics*, vol. 22, No. 1 & 2. Spring & Fall, 1994.

McGINN, Colin. *The Subjective View*, Oxford U.P., 1983.

'Can we Solve The Mind-Body problem' *Mind*, No. 98, 1989. Reprinted in Warner, 1994.

MERRICKS, Trenton. 'A New Objection to A priori Arguments for Dualism', American Philosophical Quarterly, 31, 1, January 1994. pp 81 - 84.

NAGEL, Thomas. *Mortal Questions.* Cambridge U.P., 1979.

The View from Nowhere. Oxford U.P., 1986.

'What is it like to be a Bat?', reprinted in *Mortal Questions*, pp 165-80, and Rosenthal, pp 422 - 428.

PEACOCKE, Christopher. *Sense and Content.* Clarendon. Oxford. 1983.

PEREBOOM, Derk. 'Bats, Brain Scientists and the Limitations of Introspection', Philososphy and Phenomenological research, Vol. LIV, No. 2, June 1994.

PUTNAM, H. Psychological Predicates. in Capitan and Merrill (eds), *Art, Mind and Religion.* Pittsburgh Pa: University of Pittsburg Press, 1967.

Representation and Reality. MIT, 1988.

QUINE, W.V.O., 'States of Mind', reprinted in Rosenthal, pp 287 - 8.

ROBINSON, Howard. *Matter and Sense.* Cambridge U.P. 1982.

Objections to Physicalism, Oxford, Clarendon Press. 1993. (ed.)

'Dennett on The Knowledge Argument', *Analysis*, 53, 1993a.

Perception. Routledge, 1994.

RORTY, Richard. 'Incorrigibility as the Mark of the Mental', Journal of philosophy, 67, 1970. pp 399 - 424.

'Persons Without Minds', Philosophy and The Mirror of Nature, Princeton U.P. 1979. Excerpts in Rosenthal, pp268 - 286.

ROSENTHAL. D.M. (ed.) *The Nature of Mind*. OUP, 1991.

SELLARS, W. 'Empiricism and The Philosophy of Mind'. In *Science, Perception and Reality*. London: Routledge and Kegan Paul, 1963.

'The Identity Approach to the Mind-Body Problem', Review of Metaphysics, 18, 1965. pp. 430 -51

SHAFFER, Jerome. 'Mental Events and the Brain'. *Journal of Philosophy* LX, 6, 1963. pp 160 - 66.

SHOEMAKER, Sydney. 'The Inverted Spectrum'. *Journal of Philosophy*, LXXIX, 7, 1981, pp 357 - 81.

'Lovely and Suspect Ideas', Philosophy and Phenomenological Research, Vol. LIII, No. 4, December 1993.

'Self-Knowledge and "Inner Sense"', Lectures 1 - 3, *Philosophy and Phenomenological Research*, Vol. LIV, No. 2, June 1994.

SMART, J.J.C. 'Sensations and Brain Processes'. *Philosophical Review*, LXVIII, 1959. Reprinted in Rosenthal, pp 169 - 180.

SMITH, Peter. and JONES, O.R. *The Philosophy of Mind*, Cambridge U.P., 1986.

SMULLYAN, R.M. 'An Epistemological Nightmare' in Hofstadter and Dennett, 1981.

STRAWSON, Galen. *Mental Reality*. MIT Press, 1994.

TYE, Michael. 'The Subjective Qualities of Experience'. *Mind*, 98, 1986.

The Metaphysics of Mind. Cambridge U.P., 1989.

WARNER, Richard. and SZUBKA Tadeusz (eds). *The Mind Body Problem*. Blackwell. 1994.

Includes;

WARNER, Richard. 'In Defense of a Dualism', Chapter 26.

WHITE, Stephen. 'The Curse of the Qualia', *Synthese*, vol. 68, 1986.

UNIVERSITY