MENTAL PREDICATES

Some Problems of Topic Neutrality in the Mind-Body Problem

By

Christian Edward Mortensen, B.A. (Qld.)

Thesis submitted to the University of Adelaide
for the degree of Doctor of Philosophy

Dept. of Philosophy
University of Adelaide

January, 1976

For Cathy

CONTENTS

ACKNOWLEDGEMENT

I would like to thank the following people: my supervisors, Alan Reeves and Graham Nerlich, for their help and encouragement; Michael Bradley for spending much time discussing the matters in this thesis with me; Jan Whittle for typing the thesis and Norton Ladkin for reproducing it; and, finally, my wife, Cathy, for putting up with it all.

This thesis contains no material which has been accepted for the award of any other degree or diploma in any University.

To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except when due reference is made in the text of the thesis.

Christian Edward Mortensen

SUMMARY

The thesis is in two parts. In the first part, various strategies open to the physicalist to reconcile the use of mental predicates, particularly afterimage predicates e.g. "x has a red square afterimage", with physicalism are investigated. It is argued that all the strategies open to the physicalist must leave open the possibility that what we know, or can come to know, about our minds, are facts which cannot be accommodated within a physicalist framework. In the second part, an investigation of what we can know about the state we are in when we allegedly have a red afterimage is undertaken. It is argued that we know too much for the physicalist to accommodate.

What follows is a chapter-by-chapter summary.

Part One. The Logic of Physicalism.

Chapter One.

A definition of physicalism using the idea of the set P of physicalistically acceptable predicates is given. Some properties of P are investigated.

Chapter Two.

Definitions of reduction by biconditionals and elimination by biconditionals are given. The notion of topic neutrality is explored, and two types are distinguished. It is argued that the topic neutral analysis was an attempt to show that mental predicates are members of P. Smart's original version of the topic neutral analysis is examined and rejected in favour of a more holist approach to the analysis. Two interpretations of Smart's defence to

Bradley's objections are indicated.

Chapter Three.

Lewis and Smart's later version of the topic neutral analysis, using the device of Ramsey Sentences, is examined. It is argued that this approach avoids the problem of giving an analysis of afterimage predicates which ties them too closely to a particular sort of stimulus and response. It is argued that a physicalism using this approach must still rely on Ockham's Razor and similar Principles of Method for rationality of belief in physicalism. An attempt, due to Brandt and Kim, to show that Ockham's Razor in the service of the Identity Theory is not a very strong principle is examined and rejected.

Chapter Four.

The device of adverbialisation of mental predicates in the defence of physicalism is examined. It is argued that adverbialisation without providing a semantics for the adverbialised predicates is inadequate. A parallel problem arising out of a recent debate between Davidson and Chisholm on the ontological status of events is examined. It is argued that recent work by Rennie on the semantics of adverbial contexts can provide the physicalist with what he or she needs in the way of semantics. It is then argued that it is not enough for the physicalist to give a semantics for mental predicates: the semantics must also be argued for as a true account of the world. A recent argument due to Frank Jackson against adverbial accounts of mental predicates is examined and criticised. A similar argument is tentatively suggested.

Chapter Five.

More general definitions of reduction and elimination are given. The second interpretation of Smart's defence to Bradley's objections is given: mental predicates are eliminable. Eliminative materialism is defined, and Rorty's version of it examined. It is argued that Rorty's position suitably modified can resist certain objection which have been made to it. It is argued that Rorty's position, like that of Smart and Lewis in Chapter Three, must face the test of introspective knowledge.

Chapter Six.

Another kind of eliminative materialism depending on the notion of the contingent identity of properties is defined. It is argued that there can be true contingent property identity sentences, and that there can be pairs of nonsynonymous predicates which express the same properties. It is argued that, like Lewis' and Rorty's positions, this view must also allow the possibility that too much is known about our minds to allow the identification of mental properties with physical properties.

Chapter Seven.

Functionalism is defined. Fodor's, Putnam's and Lycan's versions of it are examined and rejected.

Chapter Eight.

The possibility that the mental might be irreducible and yet physicalism be true is canvassed and rejected.

Part Two. Introspection and Perception.

Chapter Nine.

Introspective awareness is defined. It is argued that we know quite a lot about our mental states and properties. We know about their similarities, differences, causal relations and causal tendencies. It is argued that we know, or at least can come to know, more about them even than this. A class of names is invented for a class of properties of ourselves which we can know.

Chapter Ten.

Four versions of Direct Realism (as a theory of perception) are defined. Only two are taken seriously. It is argued that one version is false and another is true, but that this latter version is better described as a type of Representative Realism. In the course of Chapters Nine and Ten, it is argued that those special properties isolated in Chapter Nine as properties which we can know ourselves to have, are not identical with beliefs or suppressed tendencies to believe (i.e. belief-like items), contrary to what Direct Realism seems to be committed to.

Chapter Eleven.

A last, and formidable, argument for physicalism is investigated: that mental properties are simply identical with physical properties and science in the course of time can be expected to reveal this. Certain of our mental properties are shown to have a quasi-topological ordering. It is argued that if mental properties are identical with physical properties, then entities displaying that ordering ought to be ~~isolable~~ isolatable in, or as properties of, the brain. Some recent research in

neurophysiology is reviewed, and it is concluded that while there are items in the brain which to some extent display the same sort of topology, these items cannot be identified with the items revealed in introspection. It is concluded that physicalism is most likely false. Dualism is conceived of not as a finished doctrine, but as a research programme.

PART ONE

THE LOGIC OF PHYSICALISM

PART ONE. THE LOGIC OF PHYSICALISM

CHAPTER ONE. PHYSICALISM

1. Introduction.

In this essay we will be inquiring whether the phenomena of visual perception, afterimaging and visual hallucination can be accomodated within a physicalist world view. In Part One, we will be looking at various defences of physicalism with the aim of determining their strengths and weaknesses. In Part Two, we will investigate some of the phenomena of introspection, to see if they can be reconciled with physicalism by means of one or more of the defences of it.

The essay proceeds in the context of a fairly liberal ontology of objects, events, states, processes, and, especially, properties. As little as possible is assumed in the course of the argument about the nature of such entities; all that is supposed about properties, for example, is that they (1) exist, (2) are universals, and (3) are whatever makes predication possible. Some such working assumptions are necessary in any work which is closely concerned with classes of predicates and their relation to the world. Naturally, in another context I should be prepared to argue for my view of predication. Someone who rejects the above assumption can, I think, fairly be asked for their account of the nature of predication. It is my view, here unargued for, that many of the arguments in this thesis can be adapted to be used within a broadly nominalist ontology.

With this preamble, let us ask: what is physicalism?

2. The Ideal of a Physicalist Language.

If something is physical, it has only physical properties. If all the properties of a thing are physical, then the thing is physical. Physical things are capable of causal relationships with other physical things. These ideas are all part of what seems to me an intuitively very plausible idea of what it is to be physical.

If everything is physical and the only instantiated properties are physical, it ought to be possible (in principle, anyway) to have a language, the predicates of which denote physical properties and the constants of which refer to physical things (including, if necessary, objects, events, states and even perhaps physical properties); and to be able, using this language, to state any fact about which things exist and which properties and relations they have. We might term the belief in such a possibility, the Ideal of a Physicalist Language.

Imagine we have a suitable store of predicates, to be thought of as predicates the instantiation of which is not objectionable to physicalists. We will not say just yet what determines the make-up of such a set of predicates; we will just refer to the set P, as the set of physicalistically acceptable predicates. We will say that a physical theory is a theory (of some order, not necessarily first, but we will generally make the simplifying assumption that we are working with first-order theories) all of whose predicates are members of P. Then we can say that physicalism is the doctrine that the true and complete theory of the universe, the

true theory that leaves nothing out, is a physical theory. We will also call this theory the final theory.

There is an apparent difficulty about speaking of the final theory of the universe. In fact there are two difficulties. The first is that there is no reason to believe that there is such a theory. The second is that there is no reason to believe that anyone will ever discover it.

In support of the first point, an analogy with Peano arithmetic might be drawn. Following the work of Gödel and Rosser, it is known that there is no recursive set of axioms for arithmetic which will capture all of arithmetic. For any such recursive set of axioms, there is a sentence of arithmetic called the Gödel sentence which, if the axioms are consistent, is both true and unprovable from those axioms. So not every fact about arithmetic can be captured by arithmetical theories (of a certain sort). Further, it can be shown that if the set of axioms is strengthened by adding the Gödel sentence (so that it now becomes provable from a recursive set of axioms), another sentence can be generated which is unprovable but true if the axioms are consistent. So there is a sense in which arithmetic could never be completed, at least by the process of trying to give it a recursive basis. And physics for example might be incomplete like this. Indeed the fact that present day (and presumably future) physics requires the mathematics of the continuum, which is stronger than the mathematics of natural numbers, would seem to make this possibility even likely.

In support of the second point, it might be said that human abilities for storing and dealing with

information are relatively limited, and it is a very big universe. It might be for all we know that the problem of the final theory is too hard for humans or for that matter for any other sentient race which evolves. This point bears on the first point too, because if the final theory is never discovered, we will have to say that it somehow exists anyway; but then, if a theory is a set of sentences, the final theory will be identical with the null set. But then there will be no way of distinguishing a physical final theory from one which is not wholly physicalistically acceptable.

There are three ways (at least) of dealing with these problems concerning the definition of physicalism. The first involves the notion of limits. We might say that if we cannot hope to discover the final theory, we can at least hope to approach it. If we believe that science makes <u>progress</u> in describing an objective mind- and language- independent reality, then we might believe that successive attempts at the truth, <u>in the long run</u>, more nearly approximate it. So in the long run you would expect the non-physical bits to be removed from our theories if the world is physical and if our theories describe it more and more the way it really is. The situation is rather like the theory of limits in classical analysis. To approach a limit continuously is to be such that given any distance from the limit, no matter how small, there comes a stage after which the distance from the limit is always less than the given distance. So we might define physicalism as the doctrine that the limit of theories about the universe is a physical theory. However, this attempt turns on there being a suitable measure of distance from the truth, for theories. A candidate for this measure might be, for example, Popper's notion of

verisimilitude. Recent work by Miller,[1] however, has put this notion under a cloud. The amended definition of physicalism, therefore, awaits a clarified notion of distance from the truth, and there is reason to doubt that such a clarification might ever be forthcoming.

The second way of dealing with these problems about the final theory and physicalism is to suppose that our final theory is a set of propositions, not sentences, with identity conditions which enable such sets to be distinguished even though the propositions are never exemplified by sentences. This alternative has the problems that go with the notion of propositions and their identity conditions. The third alternative, and the one we will take, is to make the simplifying assumption that the final theory will one day exist. (We will use this device of making simplifying assumptions at other points in this thesis, too.) We do not, I suggest, know for sure that the final theory is unobtainable. It seems a fair bet, therefore, that the problems of this essay will be unaffected by the alternative possible conclusions that it is, or that it is not. So we might as well assume that it is obtainable, if it makes stating the problem any easier. Even if we knew that it was not, it does not mean that we could not assume for the purposes of our argument that it is, if the problems we want to deal with stay the same.

So we have on the one hand the Ideal of a Physicalist Language and our definition of physicalism, and on the other hand the intuitive idea of physicalism - the idea that everything is physical, with physical

---

[1] See e.g. Miller 1974-5.

properties and physical relations. These two notions are evidently closely related. Perhaps the best way to see the relationship is as that holding between a theory or definition, and a kind of loose pre-theoretical conception or insight. We might (here) call the latter a "paradigm". The definition is intended to "fit", that is, be more or less equivalent to, the paradigm in the area where the paradigm is reasonably clear. Where it is not, perhaps there are alternative non-equivalent definitions, and the considerations relevant to accepting one as against others are in addition to the fit with the paradigm; perhaps fit with another paradigm, unification of various paradigms, simplicity, or other theoretical considerations.

Nothing has been presupposed so far about what precisely physicalism is. The above could be a description of idealism, and the Ideal of an Idealist Language. What distinguishes physicalism from idealism, is the membership of the set P (and the corresponding idealist set). We will now look at the question of the membership of P.

3. The Class of Physicalistically Acceptable Predicates.

Early in his career[2], Carnap held that scientific terms were to be analysable into an observation language, conceived of as a sense datum language. Along with this belief went the view that terms which were not so analysable were meaningless - positivism in other words. Under pressure from Neurath and others, Carnap abandoned the details of the first part of this position,

---

[2]For an exposition of some of the matters in this section, see the entry "Carnap" in the Encyclopaedia of Philosophy. (See bibliography: Edwards 1967)

but not the spirit. Scientific terms had to be analysable into an observation language of mind-independent objects in space-time.

The whole view was called physicalism. One reason for this change was the recognition that science is essentially intersubjective; that many of the matters it deals with are matters independent of human minds. But if theoretical terms are to be analysed into observation language, or even if theories are merely to be based on observation sentences, in some (weaker) sense of "based on", then the observation language ought to be a language of intersubjective objects and events; and a sense-datum language is a language of private, subjective, mind-dependent events. In our terms, Carnap's physicalism amounts to a proposal to allow in P only predicates which apply to macroscopic spatio-temporal objects and events.

If all terms which are meaningful can be analysed in terms of this sort of P, and if psychological predicates are truly and hence meaningfully applicable to human beings, then psychological terms are analysable into the public observation language. Now behaviourism does not follow from this[3], but Carnap thought it did. He thought that "x is a pain" or "x has a pain" were analysable

---

[3] If the observation language contains "Cxy" for "x is the cause of y", and a store of predicates "B_1x", "B_2x", ..., for "x is behaviour B_1" etc., then "x is a pain" $\overset{df}{=}$ "$x = (iy)(Ez)(B_1z \,\&\, Cyz)$" is an analysis of "x is a pain" into the observation language, without any of the usual features of behaviourist analyses e.g. identification of pains with behaviour or tendencies, or reduction of pains to behaviour "without remainder".

into some complex of predicates about whatever is intersubjectively observable when x has a pain, that what is intersubjectively observable is x's behaviour, and so whatever is named in the sentence "a has a pain", is a, and a's behaviour.

We will not be discussing this sort of behaviourism in any detail. Carnap ultimately came to reject it for what seem to me to be the correct reasons: that "x has a pain", along with such other well-known-to-be-problematic predicates as "x is an electron", are not analysable into any observation language in the way that Carnap envisaged, that strong versions of the Verification Principle of meaning are untenable, and that realism about the unobservable is not such a bad thing.

Once you get rid of the notion that theoretical entities per se are a problem, it is easier to see that the real problem for physicalists lies in giving an account of which theoretical entities are acceptable.[4] Predicates like "x is a Cartesian Mind" and "x is a vital entelechy" are central examples of predicates which would not ordinarily be thought to be members of P. To say this, however, is not to deny that the relationship a predicate in a theory bears to (relatively) observable entities is important for physicalists. One fairly obvious condition, for example, is the condition that the theory in which the predicate occurs be relevant to the observations that we make of the world; that if

---

[4] And which observational entities are acceptable, for that matter. It is not intended in this essay that we be able to make a sharp distinction between the theoretical and the observational.

a theoretical entity make no conceivable difference to what we could observe with our senses and instruments, then it be unacceptable. This sort of consideration, in effect a much weakened condition of the same sort as Carnap's earlier insistence on the analysability of theoretical terms into the observation language, appears in Carnap's later work, and also in Feigl's, which is close to it. What Carnap and Feigl term "The First Thesis of Physicalism", is the proposal to regard statements as "scientifically meaningful" if they are "intersubjectively confirmable or disconfirmable".[5] As Feigl hastens to point out, "intersubjectively confirmable or disconfirmable" "should be understood in the most liberal manner. The sort of indirect testing of assertions here allowed for includes of course the testing of only partially interpreted postulate systems. It countenances as scientifically meaningful, statements about the most remote, the most intricately concealed or difficult to disentangle states of affairs. It includes statements about unique and unrepeatable occurrences, if only they are of a type that places them within the spatio-temporal-nomological net which itself has an intersubjective confirmation base."[6] Carnap, it is true, sometimes seems to make a stronger claim: that what has just been called "scientifically meaningful" (Feigl elsewhere calls it "physical$_c$"[7]) exhausts "factual meaningfulness"[8], or

---

[5]See Feigl 1963 p.247; Carnap 1963 p.883, in Schilpp 1963.

[6]Feigl 1963 p.247.

[7]Feigl 1963 p.242.

[8]Carnap 1963 p.882.

is "sufficient for expressing everything that is meaningful to me."[9] But this stronger claim is not entailed by the earlier one, and is unnecessary and undesirable (for it implies positivism) for our purposes. Feigl elsewhere formulates it thus:

> ..."physical$_1$" which is practically synonymous with "scientific", i.e. with being an essential part of the coherent and adequate descriptive and explanatory account of the spatio-temporal-causal world.[10]

It would be rash to take this as providing any more than a necessary condition for membership of P. Otherwise, if we were to take it as sufficient as well, we should rule out by fiat the possibility that there might be intelligibly dualist items which interact in a quite lawlike way with brains, thereby constituting "an essential part of the coherent and adequate descriptive and explanatory account of the spatio-temporal-causal world." Moreover, the existence and properties of such entities might be supposed to be subject to the sort of "indirect" testing which includes the testing of only partially interpreted postulate systems, the sort which places them within the spatio-temporal-nomological net which itself has an intersubjective confirmation base. At the very least, we should not rule out by logic the possibility that there be some such entities and that they be non-physical. All that is needed is that their interaction with the physical world be lawlike for them to be acceptable according to the First Thesis.

---

[9] Carnap 1963 p.883.

[10] Feigl 1967 p.10. Elsewhere (p.57) he gives a definition of "physical$_1$" essentially the same as the above definition of "physical$_c$".

So we need stronger conditions on the membership of P, and a natural one to think of is that things are physical, and predicates true of them are members of P, if they have something to do with physics, if they fall within the subject matter of physics.[11] Oppenheim and Putnam advocate what they call the thesis of the Unity of Science, by which they mean the thesis that all science is reducible to microphysics, at least in principle.[12] Feigl puts a similar thesis thus: "the facts and laws of the natural and the social sciences can all be derived - at least in principle - from the theoretical assumptions of physics."[13] Feigl elsewhere defines "physical$_2$" to be:

> the kinds of theoretical concepts (and statements) which are sufficient for the explanation, i.e., the deductive or probabilistic derivation, of the observation statements regarding the inorganic (lifeless) domain of nature.[14]

Many other philosophers, Carnap[15] and J.J.C.Smart[16] among them, have given similar definitions.

Two separate theses deserve distinguishing here: the thesis that the laws of all sciences can be

---

[11] A complication, arising from the fact that a physical item (e.g. a table) might have dualist predicates e.g. "being seen by me" true of it, can be answered by saying that our approach here is "holistic" - i.e. attempting to define what it is for <u>everything</u> to be physical.

[12] Oppenheim and Putnam 1958.

[13] Feigl 1963 pp.227-8.

[14] Feigl 1967 p.57.

[15] Carnap 1963 p.883.

[16] Smart 1963 pp.651-2.

derived from the laws governing inorganic processes, and the thesis that the laws of all sciences are derivable from the laws of microphysics. The second thesis clearly implies the first, but the first does not imply the second. To see this, let us suppose that dualism is true and that mental processes arise in biological structures in such a way that the behaviour of these structures is not wholly determined by the behaviour of their component parts on, say, the molecular level. Biology is not reducible to microphysics. But neither, we can imagine, are the laws of the inorganic wholly reducible to those of microphysics. It might be that consciousness arises in any structure, organic or inorganic, of sufficient complexity; and when it arises, the laws governing the behaviour of the structure are not reducible to those governing the behaviour of the microparts of the structure. So the behaviour of the organic is no different in principle from that of the inorganic, and we might suppose that a general theory of the behaviour of the inorganic is developed which allows the behaviour of the organic to be deduced from it, so the first thesis is satisfied. But in this possible world both the organic and the inorganic escape the microphysical, so that the second thesis is false.

One way of highlighting the difference is to employ the concept of emergence. We will say that a law L is emergent with respect to some set of laws S, if the subject matter of S includes the parts of the subject matter of L[17], and L cannot be deduced from

---

[17] "the parts" is a little misleading, since there are levels of parts to a thing. The parts of an object might be its cells or its molecules. So strictly, emergence ought to be relativised to some level of parts, though in practice S determines the level. We will ignore this complication. See Meehl and Sellars 1956, Nagel 1960 Ch. 11.

S together with a description of how the subject matter of L is made up of those things which are the subject matter of S. Then the first thesis is the thesis that no laws, in particular not the laws of the biological and social sciences are emergent with respect to the laws governing inorganic processes, and the second thesis is the thesis that no laws are emergent with respect to the laws of microphysics, at least for all those laws which appear in the final complete theory. Our example involved supposing that the inorganic was emergent with respect to microphysics.[18]

In terms of the set P, then, these amount to two proposals: that the membership of P be restricted either to the predicates occurring in microphysics, or to the predicates of physics and inorganic chemistry. Both of these definitions of P have problems, however.

Take, for instance, mathematical entities. It is inconceivable that future physics should be able to avoid the use of mathematics, in particular real and complex analysis. A fully developed physical theory, then, ought to contain a development of the theory of the mathematical symbolism being used in it, and it is notable that attempts to develop the theory of real numbers all require quantification over sets, or some equivalent.[19] So the final complete theory will presumably contain sentences asserting that predicates like "x has cardinality $\underline{c}$ " are instantiated; but this

---

[18]This is not such an unusual idea. In atomic physics, the Pauli Exclusion Principle was at one time emergent w.r.t. the laws governing the free motion of protons and electrons. See Feigl 1963.

[19]e.g. functors.

predicate's being a physical predicate certainly conflicts with some intuitions about the physical.

That should be
^"x has cardinality c" ~~being~~ a physical predicate might not worry a gradualist like Quine, but it conflicts with the idea that physical things are in space and time; and at least are capable of causal interaction. We might try to separate characteristically mathematical predicates of microphysics from characteristically physical predicates by some such device as requiring that members of P be true, if at all, only of objects in some region of space-time. But these sorts of moves conflict with another feature of physicalistically acceptable predicates and theories: the open endedness of P. There is nothing essentially sacrosanct about the spatial as far as physics goes. We might suppose physicists postulating and investigating particles which are not in space but which interact with particles in space. Indeed, I think that this goes for any attempt to delineate P sharply. It is not that there are some predicates acceptable to physicalists, and some not. There are degrees of acceptability, some things are more acceptable than others; and so whatever conclusion we come to about membership of P, it ought to be deliberately intended to have unsharp edges.

This does not deal with sets, though. The answer is, I think, to take a philosophic attitude to being stuck with them. If we have to say that the predicates of set theory are members of P, then, if that is all we have to deal with, things are not too bad. This is one place where we go beyond our preanalytic paradigm. Taking thelinguistic approach to capturing the physical has led us (if it does) beyond our vague

intuitions about what the world would be like if everything were physical. Of course, nothing is sacred about this paradigm, any more than any other preanalytic intuitions or linguistic forms are sacred.

Restricting membership of P to microphysics and mathematics is too restrictive, however. Not all emergence need be dualist emergence. It just might be that the final theory requires emergent principles to deal with structures more complex than the microphysical, just as for example, Pauli's Exclusion Principle was needed within microphysics without this implying that there were nonphysical principles in operation in atomic structure. Before its reduction to the Orbital Theory, the classical theory of valence was a theory which imported principles into the theory of molecular construction and interaction, which were emergent with respect to the laws governing the free motion of atoms which made up the molecules. In fact, of course, modern physics is in a very incomplete state in that there are a large number of reductions to be done which have not yet been done, and which might, therefore, conceivably not be done because they cannot be done. It would be a mistake for a physicalist to rely too heavily on the possible future reduction of everything to microphysics. Let physicalism, then, be determined by the condition that P contains only the predicates of physics and inorganic chemistry.

4. The Meanings of Predicates.

A final point remains to be cleared up before we proceed to the main task of this essay: the task of enquiring whether certain psychological predicates are members of P. The point concerns what sort of thing is

a member of P. We have been speaking of P as being a set of predicates. Now under one common interpretation of the word, "predicate" is an entirely syntactic notion.[20] Predicates are physical inscriptions or utterances, or (equivalence) classes of these things determined by the relation of same shape. But which of such things are members of P is not particularly interesting. When we speak about certain objects being members of P, we are thinking of the objects as having one meaning rather than another. The fact that in this discussion we have been speaking English might mask the fact that when we use quotation marks to name the members of P we are not thinking of P as containing inscriptions, but words of a language; and the fact that they are in a given language uniquely identifies their meaning up to ambiguity and vagueness. So one convenient way of thinking of the members of P is ordered pairs of an inscription and a meaning.

Meanings need not be particularly mysterious for our purposes. Theoretical terms typically derive meaning from the theory in which they occur, and its intended interpretation. So the second member of each of our ordered pairs might be thought of as itself an ordered set of a theory and a set of functions giving the semantics of the theory. It might well become necessary to give a semantics for varying possible worlds (in line with the idea that the meaning of a predicate is its extension in every possible world). Any sort of general decision as to the nature of the second member, however,

---

[20] Albeit perhaps nonextensional. We will not pursue the question of distinguishing predicates which have never been inscribed.

we hope to avoid. We employ the conveniently vague word "meaning". Naturally, decisions about which meanings accompany a given inscription will sometimes exercise our attention.

It might be thought that we could dispense with the syntactic items altogether and merely inquire whether certain meanings are members of P thought of as a set of meanings. This is based on the previous insight that it is meanings, not words, that are important to us. But that would be inconvenient because we should have to invent a set of words to refer to meanings, because words with quotation marks around them do not in English. And it would be unnecessary because given an understood semantics for a given predicate, using the predicate uniquely determines the meaning (up to ambiguity and vagueness): in a sense, that is what communication by means of conventional signs is all about. If we think of an interpretation of a language as a function which takes syntactic objects to meanings ~~or semantics~~, then the mathematics of functions assures us that we can use our syntactical objects to index our meanings. Reference to the index, given that the function is determined, can be regarded as implicit reference to the meaning.

To sum up then, we will be inquiring whether certain predicates with certain meanings are acceptable to physicalists; perhaps by virtue of being synonymous with some first-order function of the predicates used by physicists, perhaps by virtue of being somehow eliminable in favour of some such function without there being any previously stateable fact left unstateable. We will now turn to this question.

CHAPTER TWO. ANALYSIS

1. The Problem of Mental Predicates.

Mental predicates present a prima facie problem for physicalism. The problem is as follows. First, predicates like, "x has a red afterimage" are not, on the face of it, synonymous with any physicalistically acceptable predicate. They are not, apparently, members of P. Second, such predicates are sometimes true of things that exist, specifically people. Some people sometimes have red afterimages. But from these two premises we may conclude that if the final theory is to state every fact about which things exist and which properties they have, then it will need to contain sentences true only when certain non-members of P are instantiated. Thus, physicalism as we have defined it is false. To put it slightly differently, some things that exist have some nonphysical properties.

In this and the other chapters of Part One, we will examine various ways in which a physicalist might try to avoid this problem. In this chapter, we will look at the attempt to avoid the problem by producing an analysis of the troublesome predicates in terms of members of P. We will conclude that it is unsuccessful. In the next chapter, we will look at a more sophisticated attempt which can also be seen as an attempt at analysis. It will be useful for the discussion to introduce some terminology, and we will do that now.

2. Reduction and Elimination.

Let us consider two extensional first order theories, T_1 and T_2. The languages of T_1 and T_2, written $L(T_1)$, $L(T_2)$, are sets containing predicate constants, and any predicate or sentence which can be made up from them using the usual truth functional operators and quantification over individuals. T_1 and T_2 themselves are sets of such sentences (i.e. with no free variables) closed under the relation of logical consequence. Often in real life the membership of a theory is determined by a set of axioms, but we will not make that assumption here. We will assume that T_1 and T_2 are consistent.

Suppose that we have two n-adic predicates $\alpha \in L(T_2)$, $\beta \in L(T_1)$. The free variables of α and β can be supposed without loss of generality to be the same. Let them be "x_1", ..., "x_n". A situation we will be discussing will be called "reduction of α (in $L(T_2)$) to β (in $L(T_1)$) by means of biconditionals". This occurs when for any sentence in T_2 containing α, (1) that sentence also belongs to the closure of $T_1 \cup \{$"$(x_1)(x_2)$ ... $(x_n)(\alpha \equiv \beta)$" $\}$ under logical consequence, and (2) "(x_1)... $(x_n)(\alpha \equiv \beta)$" is true. T_2 will be said to be reduced to T_1 by means of biconditionals, when for every predicate $\alpha \in L(T_2)$, there is a predicate $\beta \in L(T_1)$ such that α is reduced to β by means of biconditionals. We will also say that a predicate $\alpha \in L(T_2)$ is eliminated (in T_1 by means of biconditionals), if T_1 reduces to T_2 by means of biconditionals, and no synonym of $\alpha \in L(T_1)$.

The point of the second clause in the definition of reduction by means of biconditionals is to ensure that the reducing theory has something to do with the

reduced theory. If we did not have any such condition as (2), then it would be sufficient for T_1 to reduce T_2, if T_1 had the same "structure" with respect to the reducing predicate as T_2 had, no matter what the subject matter of T_1 was.[1] As an example, two quite dissimilar species of plant might have a similar veinous structure, and this veinous structure be due to the presence of quite different molecules in the two plants. If the two theories about veins are genuinely reduced by the two molecular theories, so that in each of the molecular theories there exist predicates which behave in those theories in a similar way over the restricted range of application of the predicates to be reduced in the veinous theory, then without the restriction of condition (2) we should be in the anomalous situation of having to say that the molecular theory of the first species reduces to the theory of the veins in the second species. Other examples are not difficult to find.

There is a reason, too, for saying that T_1

---

[1] Alternatives to condition (2) might be either to augment it with, or replace it by, one or both of the conditions that the biconditional be "lawlike", or that it be "well-established". Apart from the desirability of avoiding the concept of lawlikeness if possible, it turns out that there are advantages connected with the notion of property identity in not insisting on the lawlikeness of the biconditionals. See below, Ch. Six. Well-establishedness is to do with the epistemological success of a theory, not its truth, and it is the latter that we are more concerned with in this essay. In particular, we are concerned with whether in certain cases there is a reduction rather than whether anyone has realized it. These are difficult points admittedly, and it is to be hoped that the difficult cases for any theory of reduction will not arise to trouble us in the sorts of cases we will be discussing. See Kemeny & Oppenheim 1956, Nagel 1960.

reduces T_2, rather than saying that T_1 U {biconditionals} reduces T_2. The latter would have the consequence that if A reduces T_2, then all the predicates of T_2 belong to L(A). This restriction would have the disadvantage that we could not consistently say both that A reduces T_2 and that a given predicate is eliminated in A. As we shall see, there are advantages in being able to say that the final theory both reduces a given theory and also eliminates a given predicate within it.

These definitions are intended to be an approximation to one type of what is more ordinarily called "reduction". It is intended that the formally defined situations above share some of the features of real life, and where they do not coincide, those places where they do not will not be interesting for our purposes and our ignoring them will not matter. For example, it follows from our definitions that any theory reduces all its subtheories[2], and that every predicate of a theory reduces itself. Our cases of reductions will usually be chosen to fit, however, more common situations. One reason for not restricting the definition so as to rule out the above cases is that we are not interested in a precise characterisation of what the restriction will be.

Another part of our simplification is that our theories are extensional and hence lack opaque contexts, e.g. belief predicates; a fact which might be thought to be troublesome if we wish to use our formalism on the problem of the status of mental predicates. Now it

---

[2] S is a subtheory of T iff T is a theory, $S \subseteq T$ and S is closed under consequences.

might be troublesome, but that remains to be seen. We will have more to say about belief and knowledge later, but in point of fact we will not principally be concerned with whether predicates like "x believes that..." are members of P; we will be almost exclusively concerned with whether visual mental predicates are. In any case, even granting the arguable point that first order theories cannot properly deal with opacity by quantifying over suitably intentional objects, more complicated definitions can always be introduced if necessary.

With these definitions, we can restate the prima facie argument against physicalism given earlier. Suppose that the troublesome mental predicate is "Mx". Now granted that we could show that there was a necessarily true biconditional linking "Mx" to some member, say "Bx", of P, there would be reason to think that "Mx" could be eliminated from the final theory, for it is a plausible sufficient condition for "Bx" and "Mx" to express the same property, that the biconditional "(x)(Mx ≡ Bx)" be necessarily true. But even if we could find such a "Bx" in P to which "Mx" might be reduced by biconditionals, if it is merely contingently true, then "Mx" cannot be eliminated from the final theory. If "(x)(Mx ≡ Bx)" is contingent, then "Mx" is not synonymous with "Bx". If "Mx" is not synonymous with "Bx", then the property of being M is not the same as the property of being B. Therefore, even if "(Ex)Mx" is materially equivalent to "(Ex)Bx", what makes them true are different facts, different properties are instantiated. Therefore, if "(Ex)Mx" is true, a language which does not contain "Mx" or any synonymous expression cannot hope to be the language of the final theory.

If mental predicates are not synonymous with any predicates from neurophysiology or behavioural psychology, the physicalist would seem to need to eliminate them from the final theory. A prima facie way of doing this is to reduce them with biconditionals and the obvious candidates for reducing predicates are neurophysiological ones. One reason, perhaps the principal one, for neurophysiology being the candidate, is that mental predicates play an important role in causal explanations of human behaviour, and it is neurophysiology which appears to be the most hopeful candidate for a unified causal explanation of human behaviour and bodily movement. It must be granted, of course, that neurophysiology is a far cry from physics and inorganic chemistry. However, it is not such an unreasonable expectation that neurophysiological predicates are reducible to strictly physical ones. There is some doubt about this question, to be sure, as will be indicated in connection with recent work by Putnam and Davidson[3]. But even if we allow this doubt to be a reality, we must still surely agree that strictly neurophysiological predicates e.g. "x is a cell" are acceptable to physicalists, and from now on we will be taking it that such predicates are acceptable.

There do seem to be cases of contingently true biconditionals where the properties associated with the predicates on either side of the biconditionals are not the same, e.g. "(x)(x is a creature with a heart ≡ x is a creature with lungs)". On the other hand, many philosophers have claimed that there can be contingent

---

[3]See Chapter Eight.

identity of properties[4]. In fact, we will be arguing that this is the case. If it is the case, then the move in the prima facie argument against physicalism from: "Mx" is not synonymous with "Bx", to: the property of being M is not identical with the property of being B, is dubious. At the moment, however, we are concerned not so much with whether there can be contingent identity of properties in reduction by biconditionals, but with what can be done for physicalism by way of analysis.

3. Analysis of Mental Predicates.

Now at first sight it looks as if well known examples which seem to fit our account of contingent reduction reasonably well, such as temperature and mean kinetic energy, genes and DNA, are also cases where the reduced predicate is eliminated or could be eliminated with no loss of ability to state which things exist and which properties are instantiated. Closer examination, however, makes it look more as if those cases where expressive power has not been lost are cases where in fact no elimination has taken place.

Consider for instance the still-to-be-achieved, but hoped for, reduction of classical genetics to molecular genetics. Now certainly "(x)(x is a gene ≡ x is a DNA molecule)" is (part of) a contingent reduction of genes to DNA molecules. But surely, also, the predicate

---

[4]As we have set it up, our immediate problem is not about the identity of properties, but the eliminability of predicates. This is an important independent consideration because with quantification over properties (essential for stating their identities) it might turn out that the properties have ineliminable and unacceptable predicates true of them.

"x is a gene" is as much a part of the reducing theory as the reduced one. Genes were initially supposed to be those entities, whatever they were, that were responsible for the transmission of hereditary characteristics. The word "gene" meant "entity of the sort causally responsible for heredity", and the concept of an hereditary characteristic remains in molecular genetics. So then, far from being eliminated from the reducing theory, some synonym of "x is a gene" occurs in it.

This is an oversimplified account of the meaning of the above term, as we will see when we deal with Ramsey sentences. For the meantime it will do as a working assumption. Certainly, if it is correct then what follows is amply illustrated by it.

The example just cited provides a device which might be used by someone wanting to show that a certain predicate is acceptable to the physicalist. The device is to find a definition of the predicate in the language of the reducing theory, presupposed to be acceptable to the physicalist. Which reducing theory? Is it not the case that the physicalist is faced with the situation where for all he or she knows one of several reducing theories, including dualist theories, might be true? This might indeed be so, so the definition should be in terms common to the envisaged available reducing theories.

It was Smart who first used this device in connection with the problem of deciding whether the predicate "x has a yellow afterimage" was acceptable to the physicalist[5]. He offered an account of its meaning

---

[5]Smart 1959.

as follows: "something is going on in x like what goes on in x when x's eyes are open, the light is good, x's eyes are good, ..., and there is a ripe lemon in front of x's eyes."

Smart wished to identify the havings of after-images with brain states. Another way of saying this, is to say that he wanted to reduce the havings of after-images to brain states.[6] But as Smart saw, that reduction would have to be contingent. So to avoid the problem of contingent reductions raised at the beginning of this chapter, he proposed an analysis to show that the troublesome predicate was already a predicate which could with reasonable charity be said to be in the expected reducing theory, neurophysiology, but in any case was physicalistically acceptable.

Smart described his analysis as "topic neutral". To say that a predicate is topic neutral was to say that it reflected the contingency of the proposed reduction. The troublesome predicate must be <u>permitted</u> reduction to the hoped-for predicate of neurophysiology, but as well it must not <u>decide</u> the issue in favour of some neurophysiology as the reducing theory. Consistent with their

---

[6] Every biconditional whose predicates are instantiated implies an identity statement, for "$(x)(Mx \equiv Bx)$" is logically equivalent to "$(x)(Mx \rightarrow (Ey)(By \ \& \ x = y)) \ \& \ (x)(Bx \rightarrow (Ey)(My \ \& \ x = y))$", and for every identity statement "a = b", there is a biconditional "$(x)(x = a \equiv x = b)$" logically equivalent to it. So we will often speak rather indiscriminately of reductions and identifications. In general, though, our problem is about reductions rather than identifications because, as indicated earlier, an identification of Ms with Bs leaves open the possibility that there might be unreduced predicates true of things referred to by terms on either side of the identity.

having the meanings they do, according to Smart, the troublesome predicates might be reduced to neurophysiological predicates, or to dualist predicates. In terms of identity, entities like the havings of afterimages might consistently be identical with states of the cortex, or states of some dualist entity or stuff. To say that the predicate is topic neutral is to say that as a matter of logical consistency either of these identifications is possible.

This can be made somewhat more precise. The discussion might be helped by introducing the notion of a <u>set of candidate reducing theories</u>. Suppose that T_2 is the theory to be reduced, perhaps because it contains troublesome predicates, and we suppose that it is true. It has not yet been reduced to any theory of a certain sort, but there are a number of candidates around. Obviously, no theory known to be false at a given time counts as a candidate. Equally obviously any candidate augmented by the appropriate biconditionals must entail T_2, otherwise there is no reduction. There may be other considerations determining membership of a particular set of candidates e.g. that the theories have as their subject matter objects on a certain "level", say the cellular level as opposed to the microphysical level; but this might be complicated by the fact that some members of a given candidate set involve emergence, and perhaps irreducible entities and laws on a number of levels. We will leave the question of what gives people the idea that a certain theory is a candidate open. Clearly there might be no candidates known at a given time (the set might be empty), and clearly also, membership of such a set will change over time as science makes progress in squaring its theories with

evidence outside the reach of T_2.

The situation with neurophysiology and its rivals for explanation of human and animal behaviour and psychology, is that research is not far enough advanced for people to know exactly what any of the candidates are. The task of identifying the functional characteristics of neural structures is in its infancy; the "logic design" of the brain and its outputs is still pretty much unknown. All that can be said about the candidates, then, is the sort of theories they will be: neurophysiological with roughly these characteristics, or emergent, or whatever. Correspondingly, all that can be expected in respect of neutrality, is neutrality with respect to theories of certain sorts. One of these, for Smart, is the Identity Theory, (hereafter - sometimes - referred to by "IT"), which is not really a candidate theory for the reduction of psychology as we have been using the term, but a restriction on those theories Smart will admit. It is a type of theory, if you like. One imagines that Smart would not care if the choice came down to an Identity theory with this logic design for the brain as against that design. It is important to see, however, that whereas physicalism can be satisfied by identifying the bearers of mental predicates with just some entities having only physical properties, the identity theory wishes to identify (certain of) the bearers of mental predicates with only certain entities, specifically the bearers of a range of predicates from neurophysiology true only of items (states, events, processes, properties) in the central nervous system[7]. Thus a crude analytical

---

[7] Needless to say, Identity theorists do not wish to identify every bearer of a mental predicate with states of the central nervous system. e.g. the values of the "x" in "x has a red afterimage".

behaviourism identifying the bearers of mental predicates with behaviour ("a pain is a groan") is physicalist but not an Identity theory.

We should distinguish between two sorts of topic neutrality. The first is where the predicate could occur either in a physicalist reducing theory, or in a non-physicalist reducing theory. The second is where the predicate could occur either in the Identity Theory, or in some of its dualist rivals, i.e. the predicate could consistently be reduced to one from a range of predicates available for it in (various) Identity theories, and consistently be reduced to some dualist predicate. Neither neutrality implies the other, since physicalism does not imply the Identity Theory, and vice versa. As far as I can determine, Smart does not make this distinction anywhere; he runs the two together. Here is a typical quotation:

> This makes our reports of immediate experience quite open or "topic neutral" to use a phrase of Ryle's. They do not commit us either to materialism or to dualism, but they are quite compatible with the hypothesis I wish to assert: that the internal goings on in question are brain processes.[8]

Smart is both an Identity theorist and a physicalist. An Identity theorist need not be a physicalist. There are at least four ways in which a person might hold IT and deny physicalism. First, one might believe that the identity theory is true but that there exist some non-physical particulars elsewhere in the universe than in human minds e.g. gods or poltergeists. Second, one could hold that both terms of the identity, along with everything else that exists, are non-physical. Idealism, for instance, might be true. Third, one might hold

---

[8] Smart 1963b p656.

that more universals exist than one could plausibly call physical. We pointed out in Chapter One that if certain mathematical predicates were needed for physics, physicalists had little choice but to count them acceptable. But a person might believe in the existence of mathematical entities unnecessary for physics. An example of mathematical entities the need for which it is hard to believe physics will ever have, are very large cardinals, e.g. inaccessible numbers. Fourth, one might be a "property dualist". Nothing formally prevents the biconditional reduction of one member of P to another, while some non-member of P remains true of the objects satisfying the predicates. Mental states might be physical brain states, but those brain states which are mental have some irreducibly psychic properties.

Because Smart is both a physicalist and an Identity theorist, and because neither physicalism nor the Identity Theory are true as a matter of logical necessity, Smart must want his analysis to be neutral in both the above ways. We can now see why it is that the topic neutral analysis in fact shows that the predicate in question is a member of P. This is because if the predicate is to be able to occur in both a physicalist theory and a non-physicalist theory, i.e. to be topic neutral in the first way, then it can occur in a physicalist theory. But a physicalist theory is one all of whose predicates are members of P. Put slightly differently, no predicate could be topic neutral between all theories in the candidate set if some of those theories were physicalist and some purely idealist; and the only way the predicate can occur in all the members of a candidate set containing both physicalist and dualist (or more generally pluralist) theories, is if the predicate is

physicalistically acceptable.

So, for a predicate to be topic neutral in both ways, it must be open to a <u>two-stage</u> reduction. The first stage is the above topic neutral (in respect of physicalism vs non-physicalism) analysis. It is a reduction of the predicate to some member of P, and it is an analytic reduction, for the reduction in fact exhibits the meaning of the predicate. Put slightly differently, an analysis topic neutral in our first sense is a demonstration that the predicate is a member of P, and can occur in various physical theories. If it does occur in those theories, it is reducible to some predicate in those theories, namely itself (or, rather, some synonym). This is covered by our definition of reduction, since if a theory is closed under deducability, it includes all instances of "$(x)(\phi x \equiv \phi x)$" in which what replaces "ϕ " belongs to the language of the theory.

The second stage reduction is the contingent reduction of the predicates to neurophysiological predicates, and it is the second sort of topic neutrality that makes this logically possible. It is clear that this second reduction is an essential part of IT, but not essential for physicalism. If physicalism is true, then the bearers of the predicates will be identical with something physical, (for instance, dispositions to behave) but not necessarily anything in, or a state of, the brain.

Now neutrality in either sense does not <u>guarantee</u> physicalism or IT. Showing that the predicates can occur in physical theories does not show that they do occur in any of the available candidates. Furthermore, even if the predicates do occur in the available candidates, this does not mean that at some future time

the candidate set might not have changed so that the available physicalist theories are inconsistent, say, with the assumption that the mental predicates are instantiated. Similarly, even if it is reasonable to include at the present time some Identity theories in our candidate set, the passage of time might change that set so that any Identity theory has consequences inconsistent with the reduction of mental predicates to those to which they would have to be reducible if IT were to be true. For example, Smart's analysis implies that if people do have afterimages, then something must be going on in them like what goes on in certain other circumstances. And it might just turn out (for it is consistent with what we now know of the brain) that nothing interestingly similar occurs when we afterimage to what goes on in those other circumstances.

For that reason, physicalists and Identity theorists have typically relied on Ockham's Razor or some similar principle. The fact is that we do not know for sure today what any of the reducing theories are, whether there are any physicalist ones that are sufficient to do the job, and whether, if there are, any are true. On the other hand, it seems a fair bet (though one which will be questioned in this thesis) that some physicalist theory of the brain's workings is true and sufficient to account for the operations of mind. The reasonableness of this bet is something which is guaranteed by principles of scientific method like simplicity of laws, simplicity of entities, explanatory power, and so on. We can for convenience group these under the heading of Principles of Method.

Later we will be looking at some discussions which challenge the above use specifically of Ockham's

Razor as a Principle of Method. One of them attempts to deny that Ockham's Razor is necessary to make belief in IT reasonable. Another attempts to deny that Ockham's Razor makes belief in IT reasonable. At the moment, though, we will look at whether a first stage analysis of Smart's sort can in fact be made. One quick point before we begin: our principal concern is to attempt to reconcile mental predicates with physicalism, rather than IT. It is not so damaging for a scientific world view if IT is false provided that physicalism is true. So the sort of neutrality mainly interesting to us is the first sort: showing that mental predicates are members of P. Henceforth unless otherwise specified it will be that sort that will be meant by the term.

It might be thought that it would be sufficient for the physicalist's purposes if all mental predicates could be shown to be topic neutral in the first sense. For then no mental predicate could pose the threat to physicalism outlined at the beginning of this chapter. That is true, but it is important to see that the success of such a program of analysis would not establish physicalism by itself, because it could not show that the final reducing theory must be physicalist. That further aim could only be achieved if the analysis showed that the predicates could only occur in a physicalist theory, and that is not what topic neutrality establishes. It is the Principles of Method which make it reasonable to believe that the set of candidates will eventually narrow down to just one physicalist theory. The physicalist could rest reasonably content if the analytical program were successful, but not without the aid of those Principles.

4. Smart's Analysis.

Smart's analysis has received considerable criticism e.g. by Shaffer, Cornman, and M.C.Bradley[9]. I wish to discuss Bradley's criticism. I will do so for two reasons. First, because I believe it is successful. Second, because the criticism of physicalism given in this thesis draws on it to some extent, and comprehension should therefore be aided.

Smart gives two different versions of the analysis in "Sensations and Brain Processes" and Philosophy and Scientific Realism.[10,11] The first is exemplified by the analysis of "x has a yellow afterimage" as "Something is going on in x like what goes on in x when x's eyes are open, ..., and there is a ripe lemon in front of x's eyes." The second is exemplified by the analysis of the same predicate as "Something is going on in x like what goes on in x when x's eyes are open, ..., and there is a yellow object in front of x's eyes." The difference is easy to see, but not so easy to characterise, since in both cases the characteristic stimuli are identified by some predicate's being true of them. Perhaps the best we can do is to say that difference lies in the degree to which it is obvious that the analysans is not in fact synonymous with the analysandum. It is clear that the first analysis does not work. There need never have been ripe lemons, no-one need have seen one, and moreover, even if there had been ripe lemons, they need not have been yellow.

---

[9] Shaffer 1963, Cornman 1962, Bradley 1963, 1964.

[10] In fact I can find no place in his writings where he explicitly distinguishes them.

[11] Smart 1959, 1963a.

All these things could have been the case and yet people have had yellow afterimages, and used and understood "x has a yellow afterimage". The point might be put by saying that a defect of the analysis is that it ties having yellow afterimages to a particular sort of stimulus, and any such stimulus is such that it only contingently has that property necessary to produce experiences of the right sort.

The defect is remedied in the second analysis by replacing "ripe lemon" by "yellow object". In doing so, however, we obtain an analysans which is not so obviously topic neutral, for how are we to analyse "yellow"? It is this problem which Bradley's argument exploits.

Smart offered an account, in "Sensations and Brain Processes", of what it is to be yellow which he hoped would preserve the topic neutral character of his analysis. The account turned on the fact that one basis on which we sort objects into groups of "same" and "different", is their colour. This might be put another way around by saying that, among other of their properties, it is their colours that enable us to discriminate objects from one another; that is, it is the colours of objects that are causally responsible for our acquiring the abilities to perform <u>certain</u> discriminatory behaviour towards those objects. Smart, then, <u>defines</u> the colour yellow as that property of objects responsible for the acquisition of a certain class of discriminatory abilities[12].

---

[12]Smart uses "power" where we have said "property". While the two concepts are different, nothing that I can see hangs on the difference for our purposes. "Property" is slightly preferable in my opinion so as to avoid having real powers as well as properties in one's ontology. We ignore, also, a refinement in Smart's account employing the notion of a "normal observer".

The abilities are abilities to perform certain pieces of behaviour, in this case sorting behaviour i.e. placing objects into classes of "same" and "different" according to colour.

This version of the analysis is, within reasonable charity, indeed topic neutral: all we have is similarities, causes, objects and sorting behaviour. But Bradley argues that the account of colour is inadequate. His argument is that our abilities to sort objects into same colour and different colour could remain undiminished through certain changes in the colour of objects. He gives two cases. We could imagine that instead of being coloured, everything was various shades of grey, but objects which are now the same colour would in the new world be the same shade of grey, and objects which are now different colours would in the alternative world be different shades of grey. Now we can discriminate shades of grey, some people can do it very well. So our abilities to sort objects as "same" and "different" can easily be imagined to remain undiminished. But in this possible world nothing would be the colour yellow. Being yellow, therefore, cannot consist merely in having the power to cause discriminatory abilities. The second case is that we could imagine a universal colour change, with everything now red being orange, everything orange being yellow, etc. In this circumstance we would sort the same objects into exactly the same groups according to similarity and difference. Colours, therefore, are not exhausted by their abilities to cause discriminations in humans.

Since this "topic neutral" account of colour is unsuccessful, Bradley argues, we must look for another account of colour is we are to be convinced that the word

"yellow" in the analysans is topic neutral. However, the obvious account - being yellow consists not in having the power is cause discriminatory abilities but in having the power to cause experiences e.g. sensations of yellow - is no help to the materialist. This is because when this definition of "yellow" is put into the RHS of the analysis, we have the phrase "sensations of yellow" occurring on the RHS, and this is far from obviously topic neutral. In fact the analysis of a similar predicate is circular, because presumably we would not just analyse "x has a yellow afterimage" in this way, but also "x has a sensation of yellow" (provided we allow that "like" be construed so that a thing can be like itself). But then the analysis of "x has a sensation of yellow" would have that phrase itself in the analysans. This is not an objection that analyses cannot be circular; it is an objection that a circular analysis fails as a demonstration of topic neutrality.

Smart's defences against these objections are at first glance very weak. In Philosophy and Scientific Realism he concedes that Bradley's arguments show that one cannot give an account of colour concepts in the way he thought (although he also claims that inner experiences are of "very little" importance to colour concepts, and that their "inner core" is analysable in terms of discriminatory responses).[13] He says, however, that he will later in the book solve this problem by showing "that I do not need to ascribe to them (i.e. inner experiences - C.M.) any qualia or properties which cannot be dealt with in a physicalist way." (p.83) When he later comes to this

---

[13] Smart 1963a pp.82-3.

issue (Chap. 5) he merely gives variants of both the kinds of analysis we distinguished before, and says nothing to clear himself of the charge of circularity except to deny that the topic neutral formula is a "translation". "It is rather meant to give in an informal way what a sensation report purports to be about." (p.96) Later, on the other hand, he continues to speak in ways which suggest that he thinks he has provided an analysis e.g. "But the specifically neurophysiological properties are not mentioned in the sensation report, which is 'open' or 'topic neutral'" (p.97) "... our reports of inner experiences are topic neutral in a certain way: that they report inner experiences essentially as 'like what goes on in me when ...' " (p.103).

If the topic neutral formula is not a translation, it cannot constitute a demonstration that the troublesome predicates are topic neutral and hence members of P. I propose to claim, nevertheless, that there are in what Smart says the seeds of two approaches that might be taken to avoid this criticism.[14]

I will mention the first approach, but defer discussion of it until later. The words "give in an informal way what a sensation report purports to be about" might be understood to be saying that the analysans gives what we are introspectively aware of. It might be that the topic neutral "analysis" is not so much an analysis as merely a substantial claim about what is introspected. (I say "merely", because of course if the

---

[14] I do not mean to say that these are approaches which Smart was clear about at the time, only that they are approaches which might be adopted to avoid this criticism.

"analysis" is an analysis, then we shall be committed to this substantial claim too, if we want to hold that people do have afterimages.) This sort of interpretation would make Smart closer to Rorty (whom we will discuss later) than we have hitherto allowed. The claim would be that all that happens when we have yellow afterimages is that something goes on like what goes on when ...; and, we might add, if "x has a yellow afterimage" cannot be analysed in these terms, we can always deny that anyone ever afterimages.

The trouble with interpreting Smart this way is that it is clearly his view that mental predicates are true of people. He seems therefore to be in the difficult position of producing a convincingly topic neutral analysis which avoids Bradley's argument.

Another way of interpreting the intended looseness of the relation between analysans and analysandum, is to say that the RHS is not right, but that something of that sort is right; something of that sort in that it has only similarities, causal relations, and (as later, p.103) things like waxing and waning in it.

Now if this something were a short formula like the original, with just some variation in the stimulus cited, and perhaps the addition of some typical responses (following Armstrong), Smart ought to have produced it if we are to be convinced that it is what is right.[15] It might be, though, that producing the short formula is somehow impossible (though not in a way that is bad for the

---

[15] It might be that Smart had various short formulae in mind at the time, thinking perhaps that the short formulae might vary from person to person. That will not help, of course, because Bradley's problem will arise separately for each individual.

formula, as for example in the way that the impossibility of producing analyses of physical object sentences into sense-data terminology counts against analytical phenomenalism.).

After all, the original topic neutral analysis suggests that when we learn mental concepts we somehow have a characteristic stimulus-response situation in which the mental state is produced, and we are taught to apply the mental predicate to whatever mental state is produced in that situation. This is evidently an oversimplification of the learning situation. We learn to recognise our mental states and learn our mental concepts together, and we learn both as much by their relations (causal, similarity and other relations) to other mental states, as by their relations to external causes and effects. Thus, mental vocabularies are typically acquired as a whole, and acquiring mental concepts is simultaneously learning a series of (loose) laws and generalisations about mental states, both in ourselves and others. If, then, we wish to tie the concept of a mental state to the learning situation, we will have to say that the meaning of a mental term is given by the position it occupies in a theory, and the theory in question will be the sort of loose set of beliefs that we all have about our mental states. We might call it "common sense psychology".

This, then, is the second way out for Smart. It is of course the view of the function of mental language expressed in the Lewis-Smart Ramsey sentence approach which we will discuss soon.

It gives backing to the idea that the old analysans and analysandum are connected, but loosely. The connection was loose because any particular stimulus, if

connected to the mental state at all, is connected by loose, probabilistic laws (and not necessarily straightforward statements of probability, but statements of likelihood, tendency, etc.). The connection, though, is in that the analysans is the right sort of thing. We will see in the next chapter that Smart's version of Lewis' Ramsey sentence for common sense psychology is intended to have just those sorts of relations in it that the old analysis has in it.

The same points can be made about Armstrong-style analyses. Armstrong proposes to analyse mental predicates in terms of effects. The schema for analysis is "x has M" $\overset{df}{=}$ "x is in that state which typically causes behaviour B". But it is clear that many such states typically cause only other mental states, and many do not have any typical effects at all. An analytical reduction, therefore, must if it is to be successful at all, link mental predicates to a whole theory, and it is the whole theory that must be reduced "in one go" as it were.

We have seen, then, that the original version of the topic neutral analysis fails in its attempt to give a predicate-by-predicate analysis which shows each one individually to have analytical links with topic neutral predicates. The defence to physicalism by analysis must take a more holistic approach if it is to succeed, and in the next chapter we will look at this.

CHAPTER THREE. RAMSEY SENTENCES AND OCKHAM'S RAZOR

1. Lewis and Ramsey Sentences.

Smart's original version of the Identity Theory rested on two platforms: the topic neutral analysis and Ockham's Razor. It has recently been suggested by Smart[1], following work by Lewis[2], that Ockham's Razor might be unnecessary for rational belief in IT. In this and the following sections, we will discuss this claim.

The central device employed by Lewis is the Ramsey Sentence.[3] Ramsey treats theories a little differently from the way we have been. Theories are thought of as conjunctions of all the sentences in them. The conjunction might be infinite, although if the theory is conveniently finitely axiomatisable, it can be replaced by the conjunction of some finite set of axioms.[4] Let T be a first order theory all of whose constants are individual constants, and let them be "t_1," ..., "t_n," "o_1", ..., "o_m".[5]

---

[1] Smart 1970-71 p. 351.

[2] Lewis 1970a, also 1966, 1970b.

[3] After its inventor, Frank Plumpton Ramsey.

[4] In my view the approach we have been adopting is superior in that it avoids dealing with the theory of languages with sentences of infinite length, in favour of infinite sets of sentences, which are standardly treated in the model theory of first order theories. We will remain with the Ramsey-Lewis formulation, however, if only to avoid the problems of conversion.

[5] Conversion of theories with only predicate constants to theories with only individual constants, and vice versa, is straightforward. As previously indicated, formulation in terms of predicates has some advantages.

The constants are thought of as being in two classes: the T-terms {"t_1", ..., "t_n"} and the O-terms {"o_1", ..., "o_m"}. T-terms and O-terms are not intended by Lewis to be "theoretical" and "observational" terms respectively. Rather, O-terms are thought of as terms the meaning of which has been fixed independently of the theory T, and T-terms are terms that are hitherto undefined. T is thought of as introducing "t_1", ..., "t_n"; that is, T defines "t_1", ..., "t_n" as standing for those entities whatever they are which behave in the way T says they behave. To make the special position of "t_1", ..., "t_n" in T clear, T will sometimes be written $T(t_1, \ldots, t_n)$.

"t_1", ..., "t_n" are constants and so occupy unbound places in T. Let $(Ex_1, \ldots, x_n)\ T\ (x_1, \ldots, x_n)$ be the result of replacing each t_i wherever it occurs in T by "x_i", and binding "x_i" by an existential quantifier with the whole of T as its scope. We say that $(Ex_1, \ldots, x_n)\ T(x_1, \ldots, x_n)$ is the Ramsey Sentence for T, and write it $(E\underline{x})T\underline{x}$ for short.[6]

The Ramsey Sentence for T can be thought of as a way of eliminating the use of T-terms and yet retaining all the other consequences of T. To say this is not however to say that the use of the Ramsey Sentence somehow supports instrumentalism. Eliminating T-terms

---

[6]This requires that "x_i" not occur elsewhere in T. It will not occur free, because only the constants occupy free places, and if bound can be changed. The order of the quantifiers in $(Ex_1, \ldots, x_n)T$ is of course irrelevant. Unless otherwise specified, "$\underline{x}$" is to be read as (the ordered n-tuple) "$(x_1, \ldots, x_n)$", and "$(E\underline{x})$" as "$(Ex_1)(Ex_2) \ldots (Ex_n)$".

like this does not show that what those T-terms hitherto allegedly referred to are in any sense unreal. On the contrary, the Ramsey Sentence quantifies over what the T-terms of the unramsified theory referred to: a class of entities in general quite distinct from those which the O-terms stand for. Taking the quantifier seriously commits us to the values of the variables in the scope of the quantifiers.

Just as the Ramsey Sentence gives us a way of eliminating T - terms, so conversely it provides us with a method for introducing, or defining T-terms, insofar as a theory can be thought of as introducing or defining its theoretical terms. Lewis points out that when a scientist introduces theoretical terms along the lines described by the account of T-terms, he normally has in mind that his terms refer to some unique class of objects, and indeed we might think of the theory as containing implicitly or even explicitly the claim that it is uniquely realised.[7] So, given the Ramsey Sentence, we might offer, as definitions; for each i, $1 \leqslant i \leqslant n$

"t_i" $\overline{df}$ "$(iy_i)(Ey_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n)(x_1, \ldots, x_n)$ $(T(x_1, \ldots, x_n) \equiv . y_1 = x_1 \ \& \ \ldots \ \& \ y_i = x_i \ \& \ \ldots \& \ y_n = x_n)$"

or more simply, and in the object language $\underline{t} = (i\underline{x})T\underline{x}$.

There are two important points to note here. The first is that, for Lewis, the constants t_i and o_i are thought of as names for things of any sort, e.g. events, states, properties, sets, that the theory needs. They

---

[7] An n-tuple $(a_1, \ldots, a_n)$ is said to be a realisation of $T(x_1, \ldots, x_n)$, iff $T(a_1, \ldots, a_n)$. T is uniquely(multiply) realised in case it has exactly one (more than one) realisation.

are not restricted to being names for objects. Thus, for instance, where t_i names a property, the above definitions give us a definition of a property name.

The second point is that while some theories entail that they have a unique realisation, and some theories do not, if we are to think of the scientist using his theory to define its T-terms, then we must take it that the theory at least implicitly asserts its own unique realisation. That is a necessary condition of t_i being a name, and a necessary condition of t_i being well-defined in the above. There is a formal device for making this implicit assumption explicit - Lewis calls it the Modified Ramsey Sentence - but we will not go into this complication. We will simply take it that in the above, the theoretical terms can indeed be well-defined by the theory.

The definition of the terms by using the whole theory is the "holist" approach to analysis which was indicated at the end of the previous chapter. It seems reasonable to think that it is the best chance for any such analytical program, for the reasons indicated earlier.

The core of Lewis' suggestion is that often enough it is a straightforward discovery (perhaps a scientific discovery) that a certain n-tuple of objects is a realisation of T. Take a simple theory about a triple of people (x, y, z) involved in a murder. T contains as one of its conjuncts the sentence "x saw y give z the candlestick while the three of them were alone in the billiard room at 9.17 pm." It is easy to imagine that we just might come to discover that the triple (Plum, Peacock, Mustard) is a realisation of our theory.[8]

---

[8]This is an example of a theory which explicitly implies (i.e. entails) that it is uniquely realised.

In the case where we have made the discovery that $\underline{r} = (r_1, \ldots, r_n)$ is a realisation of $T(x_1, \ldots, x_n)$, we should have no choice but to identify $\underline{r}$ with $\underline{t}$. That $\underline{r} = \underline{t}$ follows directly from what we have discovered (that $T\underline{r}$), together with our definition of $\underline{t}$: $\underline{t} = (i\underline{x})T\underline{x}$, by the transitivity of identity. Notice that "r_1", ..., "r_n" need not be "theoretical" terms, or T-terms in any theory. They might be names for objects we are already well acquainted with. Thus, in our example, we might in the course of the final evening in the drawing room have proposed as definitions

Black = $(ix)(E!y)(E!z)$(... & x saw y give z etc. & ____)

White = $(iy)(E!x)(E!z)$(... & x saw y give z etc. & ____)

Green = $(iz)(E!x)(E!y)$(... & x saw y give z etc. & ____)

Now when we find that our old school friends Plum, Peacock and Mustard are such that ... & Plum saw Peacock give Mustard etc. & ____ , we have no option but to conclude that Plum = Black, Peacock = White, and Mustard = Green. Of course, "r_1", ..., "r_n" might be theoretical terms or T-terms in some reducing theory which has "$(E\underline{x})T\underline{x}$" as a consequence. In such a case we would be compelled to make a reduction of some of the subject matter of one theory to some of the subject matter of the other; an identification of $\underline{r}$ with $\underline{t}$.

An important point to emphasise here is that the identification of $\underline{r}$ with $\underline{t}$ might be something that the theory itself entails, without there being any "further" discovery of the sort $T(\underline{r})$. We allowed for this possibility in the previous chapter when we said that a topic neutral analysis was an analysis of a predicate (here, term) into the language of a candidate reducing theory. For if, for example, the candidate theory was a neurophysiological one, and the analysis of "t" was "the cause of B", then

the neurophysiological theory could normally be expected to include the identification of the cause of B with a suitable r from neurophysiology.

> And that is how, someday, we will infer that the mental states M_1, M_2, ..., are the neural states N_1, N_2, ...
>
> Think of common sense psychology as a term - introducing scientific theory, though one invented long before there was any such institution as professional science. Collect all the platitudes you can think of regarding the causal relations of mental states, sensory stimuli and motor responses. ...Add also all the platitudes to the effect that one mental state falls under another - "toothache is a kind of pain" and the like. Include only platitudes that are common knowledge among us - everyone knows them, everyone knows that everyone else knows them, and so on. For the meanings of our words are common knowledge, and I am going to claim that names of mental states derive their meaning from these platitudes.
>
> Form the conjunction of these platitudes; or better, form a cluster of them - a disjunction of all conjunctions of most of them. (That way it will not matter if a few are wrong.) This is the postulate of our term - introducing theory. The names of mental states are the T-terms. The O-terms used to introduce them must be sufficient for speaking of stimuli and responses, and for speaking of causal relations among these and states of unspecified nature.
>
> From the postulate, form the definition of the T-terms; it defines the mental states by reference to their causal roles, their causal relations to stimuli, responses and each other. When we learn what sort of states occupy these causal roles definitive of the mental states, we will learn what states the mental states are - exactly as we found out who X was when we found out that Plum occupied a certain role, and exactly how we found out what light was when we found that electromagnetic radiation was the phenomenon that occupied a certain role.[9]

---

[9] Lewis 1970a pp.17-19

2. Topic Neutrality.

Now what has happened to the topic neutral analysis in all this ? If mental predicates fail to be topic neutral, can Lewis reasonably expect it to be discovered that mental states are physical ? To be sure, Lewis leaves open the possibility that common sense psychology has no realisation at all, but that some "near realisation" of it is identical with some n-tuple of brain states. But suppose for instance, that common sense psychology were hopelessly dualistic. Suppose the Ramsey Sentence contained conjuncts like "Pain is nowhere in physical space and is not a state of things in physical space", or "Neither afterimages nor their constituents have mass", and so on; and not just contained (a few of) them, but was positively infested with them. In this circumstance, should not we be forced to conclude that if the Ramsey Sentence were true, no realisation of it was identical with any n-tuple of neural entities? Lewis and Smart do not deny anywhere that we would be forced to this conclusion in this circumstance.

Lewis and Smart give different accounts of what is in the Ramsey Sentence for common sense psychology. Lewis' sentence is largely causal , as should be clear from the above passage. It says that the referents of the T-terms bear causal relations to one another, although Lewis allows as well that there will be "platitudes to the effect that one mental state falls under another", and, in the case of intentional states, propositional attitudes, etc., relations between mental objects or persons, and propositions.[10] He does not say that such a sentence will

---

[10] And so it will lack some sort of topic neutrality in that it is committed to a metaphysics of propositions.

be topic neutral, but he does claim that it displays, or exhibits, or provides an account of the meaning of mental terms. Common sense psychology functions as if we had introduced mental terms as T-terms into a theory the O-terms of which were relatively gross descriptions of stimuli and responses, as is described by Wilfred Sellars' Myth of our Rylean Ancestors.[11] No-one of course ever did introduce mental terms in this way, so Sellars' "myth" is a myth; but it is a "good" myth - "our names of mental states do in fact mean just what they would mean if the myth were true."[12]

Smart also appeals to Sellars' Myth, "to illuminate the way in which reports of inner experience came into our use of language".[13] Smart also treats the Ramsey Sentence as a "definition", but he explicitly includes more predicates and relations into it than Lewis does: similarities between mental states (p.350, 354), "certain topic neutral words, such as 'waxes', 'wanes', 'gets more intense', 'comes intermittently'" (p.350), and classifications ("to say that something is an ache is to classify it (in terms of unspecified likenesses) as a sort of pain" (p.351)), as well as causal relations.

Of any alleged topic neutral analysis, we can ask: is it topic neutral? is it a correct analysis? Now Lewis does not argue that his Ramsey Sentence is topic neutral. But I think that any theory of Lewis' sort which asserted only causal relations between, and classifications of, its

---

[11] Sellars 1963.

[12] Lewis 1970a p.20.

[13] Smart 1970-71 p.352.

theoretical entities would pretty obviously be topic neutral if any theory would be. The debate between Identity theorists and dualists has traditionally taken place against a background assumption that some form of psychophysical interactionism - and so some form of psychophysical interaction - is possible. Furthermore, Lewis clearly believes that it is not a contradiction to say that the theory is realised by some set of physical entities, or by some other set of entities. Nor does Lewis actually argue that the Ramsey Sentence is a correct analysis.[14] He quotes Sellars' Myth with approval and leaves it at that, but he clearly intends it that the Ramsey Sentence give an account of the meaning of the terms in question. So, then, as to whether Lewis' Ramsey Sentence is a topic neutral analysis, it seems that (1) it is topic neutral and Lewis believes it consistent with the Identity Theory and dualism, (2) Lewis intends that it be a correct analysis.

Smart's version is a little different. Smart intends it to be topic neutral. He refers for instance to "general (topic neutral) truisms of common sense psychology" (p.354). Further, he is careful to mention explicitly only those predicates and relations which appear to have a good case for being called topic neutral (similarities, waxings and wanings, classifications, and causes and effects). But whereas Lewis speaks as if his Ramsey Sentence includes only those parts of common sense psychology which suit him, Smart seems to believe that his Ramsey Sentence is more or less all of common

---

[14] He does offer "one item of evidence" (p.20): that many philosophers have found analytical behaviourism plausible, that there is a "strong odour" of analyticity about common sense psychology, and that Sellars' Myth explains both these.

sense psychology. Witness the previous quote, and also: "Let us imagine common sense psychology Ramsified in the manner suggested by Lewis." (p.354) Lewis is the more cautious as regards the topic neutral status of the Ramsey Sentence here: it is not so obvious that "common sense psychology" manages to avoid dualist commitments. Nevertheless when Smart lists the bits of common sense psychology he wants to include, they turn out to be the sort of bits he has relied on in all his pre-Ramsey Sentence work for the topic neutral analysis, and bits which are prima facie topic neutral.

As to the further question of whether Smart's theory is really an analysis, the matter is obscured by the previously mentioned lack of caution. Granted that common sense psychology provides us with an account of the meaning of mental predicates along Ramsey lines, it is not so obvious that common sense psychology is topic neutral. On the other hand, if you stick to just those pieces of common sense psychology listed by Smart or Lewis, you have a Ramsey Sentence which might well be topic neutral, but it will need some argument to establish that you have captured fully the meanings of mental predicates.

Unlike Lewis, Smart sketches the argument. It turns out to be the argument which was implicit in his work in many places previously.[15] We learn mental terms primarily with respect to external stimuli and responses, and mental concepts are linked to such paradigmatic learning situations. Secondarily we introspectively

---

[15]Smart 1959, 1963a, 1963b.

recognise mental states as being similar or dissimilar to one another, and as being more or less intense, etc., without knowing the respects in which they are similar, or intense, etc. Since these are all the features of mental states that we are aware of, it is only terms for these features that can be analytically linked to mental terms (and this is presumably true no matter what the ordinary person believes). Terms standing for physicalistically objectionable features do not enter into the analysis, for we have no reason to believe there are such features.

We will not discuss this account here. We will give arguments bearing on it later. What is important to note is that Smart does propose his Ramsey Sentence as an analysis. So Smart intends his account to involve a topic neutral analysis, as he has usually done.[16]

3. Is Ockham's Razor Avoidable?

For a person who thinks that the Identity Theory <u>needs</u> a topic neutral analysis, what are they committed to in respect of some version of Ockham's Razor? A person who thinks that a topic neutral analysis is necessary presumably thinks that its failure results in the falsity of the Identity Theory. That is, the falsity of the Identity Theory <u>come what may</u> in the way of future scientific discoveries concerning the working of the brain. Notice that this does not mean that the failure of the topic neutral analysis implies the <u>logical</u> falsity of the Identity Theory. As we have already seen,

---

[16]Except for the brief "middle period", exemplified by his paper in Presley 1967.

common sense psychology need not by itself imply dualism in order for us to be forced to accept dualism. The claim that a certain mental state exists might, together with other premisses known or subsequently discovered to be true, imply the falsity of physicalism.

Arriving at the conclusion that common sense psychology is true but inconsistent with candidate materialisms should give us some expectations about the future course of science. If mental states were known to have certain properties causally relevant to human behaviour which brain states were known not to have, then whatever else we find about the brain's operations, it is a fair bet that we will not find that they sufficiently explain behaviour.[17] With a certain amount of optimism we could expect to find that they are positively insufficient. The success of the topic neutral analysis, however, does not guarantee that future science will not go along such a course. Just that course of science described just now as a consequence of a certain kind of dualist argument, could occur independently of the truth of its premisses. That is to say, there are imaginable circumstances which would make it reasonable to say that we had discovered that dualism is true. If we found a person with a head full of sawdust, we should need to look elsewhere for mechanisms of sufficient causal complexity for behaviour. The head does not have sawdust in it, so much is known. But it is also true that science at present does not explain all human behaviour: the head might still be made of sawdust of a more subtle sort. So the state of affairs that Lewis describes, where we come to discover that

---

[17] The dualist alternative would be a plurality of causes, singly sufficient to explain behaviour. See below, this section.

neural mechanisms play a certain causal role, is at the present time a hope. A reasonable hope, if you like. One made reasonable by the present successes in psychobiology, undoubtedly. But it could still at this late hour turn out to fail. And furthermore, Lewis concedes as much when he says that we will "one day find" that the brain is thus and so, rather than saying that we now know it.

What makes the hope reasonable? Induction from past successes perhaps, but induction in the absence of background theory and methodology is a risky business. The classical version of Ockham's Razor is a crude device for describing all the methodological moves scientists might make. Parsimony of entities is one sort of goodness a theory can have, another is simplicity of laws. Explanations should as far as possible appeal to known laws and mechanisms to the exclusion of others. This is not the classical Ockham's Razor, but like Ockham's Razor it is a methodological principle rather than a fact of nature. Like Ockham's Razor it does not take us from truth to truth with the force of logic, but establishes the rationality of accepting conclusions arrived at using it. It is quite a powerful principle. Charitably interpreted, it can be used to establish the reasonableness of Lewis' hope and with it, other things being equal, the reasonableness of belief in the Identity Theory: psychobiology has achieved successes with neural mechanisms, and we ought to believe that those mechanisms will be entirely sufficient.

Furthermore it is hard to see how Lewis could establish the reasonableness of his hope without appeal to some such principle as the above. If there is no premium on explaining future discoveries by today's mechanisms, or

if there is no requirement to avoid supposing mechanisms of an ontologically objectionable sort, then it can hardly be especially rational to do so.

Lewis' language somewhat obscures the point. This is how we will make the identification "one day", he says. Certainly. Barring a later qualification, perhaps then it will not need Ockham's Razor. But does not your present belief in the Identity Theory require Ockham's Razor or some other Principles of Method to be rational? A preliminary conclusion, then, is that some things like Ockham's Razor in that they are Principles of Method cannot be dispensed with.

There is a place where something very like Ockham's Razor itself intrudes, which is rather more obscured by Lewis' formulation of the reduction situation. If our Ramsified theory turns out to have no realisation, or multiple realisation, then, according to Lewis, we are to count it false.[18] What would the world be like if common sense psychology had a multiple realisation?[19] Well, if we restrict ourselves to the Lewis version of common sense psychology, it would be that there are at least two sets, at least partly distinct, of causes and effects - causes of responses and effects of stimuli.

One way in which this could come about is if there are two types of entities, one neural and one

---

[18] And so we must arrange for some formal treatment of sentences with "t" in them. e.g. the identity "r = t". Lewis uses a theory of descriptions due to Dana Scott 1967.

[19] We will not discuss the possibility of T's not being realised because of the complication arising from the Löwenheim Skölem theorem that a consistent denumerable theory will have a realisation in the natural numbers.

ghostly, each of which would be sufficient by itself to bring about the effects in question. However, when and only when an entity which is one of the neural realisations of the theory is produced, a corresponding psychic entity is produced. When neural entities produce one another, so do corresponding psychic entities. At the end point in the process whenever a neural entity produces a piece of behaviour, so does some psychic entity produce the same behaviour. In such circumstances, it is easy to imagine that we would not be able to detect the psychic entities; and, so far as neurophysiology is concerned, things would seem to be the way Lewis and Smart hope they will someday turn out to be.

This cries out for Ockham's Razor. The entities are undetectable, there is no reason to suppose they are there, so do not postulate them. Now why is not Lewis supposed to need Ockham's Razor to deal with this possibility? Simply because his Ramsified theory carries with it the implication of unique realisation. We could not get a definition of "t" as "ix Tx" going, unless T had a unique realisation. When scientists treat theories as frames within which to formulate definitions of theoretical terms, they must presuppose that their theories have unique realisations. And our example is an instance of multiple realisation.

Presupposing that a theory has a unique realisation does not make it that it has one, however. No doubt we are sometimes in a position to know that such a presupposition is true, as Lewis points out. But it is not plausible to think that in the present state of science or even in Smart's and Lewis' supposed future we would be in a position to know that the presupposition

of unique realisation of common sense psychology is true independently of something like Ockham's Razor. This appeal to ontological simplicity is different from the previous one, in that it is an appeal which must be made even in Lewis' utopia. Our previous point was that belief that science will turn out like that depends today on some Principlesof Method. But even in the future when we argue

$$\begin{array}{l} \underline{t} = (i\underline{x})\ T\underline{x} \\ T(\underline{r}) \\ \therefore\ \underline{r} = \underline{t} \end{array}$$

the appeal to the legitimacy of the definition of the first premise require Ockham's Razor.

These two points about Principles of Method can be put rather more succinctly with our idea of the candidate set. Set up the Ramsified common sense psychology, "$(E\underline{x})T\underline{x}$". The presupposition that T has a unique realisation, and the possibility that mental terms be well-defined by T, require Principles of Method for rational belief. Now look for candidate reducing theories for "$(E\underline{x})T\underline{x}$". At the present moment several types, including some dualisms, are consistent with the evidence. So the belief that the set will one day narrow down to one which will then imply the identification and which will be physicalist, is a belief which at the present time needs Principles of Method to support it.

We have not yet asked whether common sense psychology is consistent with physicalism. Later in this book we will be arguing that in a sense it does not matter because we can find out some facts about psychological

states which seem to be inconsistent with what physicalisms there are around. It will emerge that it is not clear whether these facts are to be included in common sense psychology as we now conceive it, because it is not clear whether up to now anybody knows (or believes) these facts. For this sort of reason, I think that deciding whether common sense psychology is topic neutral is not so easy, but also that the question can be circumvented.

Anyhow, in the remainder of this chapter we will pursue the other topic discussed so far, Ockham's Razor. R.M. Brandt and Jaegwon Kim have suggested that the only reason for preferring the Identity Theory over one of its alternatives is Ockham's Razor, and that this is not a very strong reason.[20] In each of their three articles discussed here, Brandt and Kim contend that the identity claim that the Identity Theory makes can be replaced by a weaker claim which has the same "empirical" consequences. The weaker claim they have in mind is the claim that a phenomenal event of a certain sort occurs iff a brain event of a certain sort occurs. The only advantage that IT has over this weaker principle, they argue, is Ockham's Razor, and this is not very much of an advantage. In their earlier articles, in fact, they hold that it is a positive disadvantage. In the later article this is modified, and it is the later article that we will concentrate on.

---

[20]See Brandt and Kim 1967, also Brandt 1960, Kim 1966. We will mostly discuss the former as it is the most recent of the three, and references will be to it unless otherwise specified. Kim has more recently modified his views. See Kim 1968, 1972.

4. Brandt, Kim and Ockham's Razor.

An event, for Brandt and Kim, is specified entirely by a triple (U_i, t_i, L_i), where U_i is a property, t_i a time, and L_i a location. "location" is taken conveniently broadly; in some instances it could be taken to be sufficiently determined by specifying the object which has the property. Two events (U_i, t_i, L_i), (U_j, t_j, L_j) are identical iff $U_i=U_j$, $t_i=t_j$, $L_i=L_j$. A phenomenal event is an event the first member of which is a phenomenal property. A phenomenal property is a property designated by a phenomenal predicate, which is "one which the person in whose experience the phenomenal event occurs might define for himself ostensively, to refer to the features of which he is directly aware... the instance of the property must be one of which exactly one person is directly aware."[21] A mental predicate is one "that can be fully explained only by a clause which makes some reference to phenomenal events of the person to whom the predicate is ascribed."[22] The Identity Theory is the theory that every phenomenal event is identical with some physical event.

The weaker claim to be compared with IT is what they call the Principle of Simultaneous Isomorphism (PSI), which is

> For every phenomenal property M, there is a physical property P, such that it is lawlike and true that for every x and every t an M-event (i.e. an event involving the instancing of M) occurs to x at t if and only if a P-event occurs in the body of x at t; further, distinct phenomenal properties have distinct physical correlates.

---

[21] pp. 518-9.

[22] p. 519.

[23] p. 521.

According to them, this is entailed by IT, and the statements of IT have "no more empirically verifiable content than their associated correlations".[24] There is no gain in the "'hard' factual explanatory and predictive power of the theory for its total factual content remains unchanged".[25] The sole reason for preferring the Identity Theory to the weaker PSI is Ockham's Razor, which is admittedly some reason, but "less than overpowering", "not compelling", "not a very strong reason".[26]

Brandt and Kim warn us to be wary of the philosophical commitments of one's theories. In this instance, they might be invited to take their advice seriously, for their views seem to commit them to a Humean view of causes. I will take it that a theory is Humean if it does not contain any primitive predicate "Cxy" for "x is a causal antecedent of y".[27] Now if a theory T_2 does contain such a relation, and it holds between things and Ms, or Ms and things, and a candidate theory T_1 together with "(x)(Mx ≡ Bx)" entails T_2, there is no guarantee that T_1 together with PSI will also entail T_2. For a non-Humean, that an M invariably and in a lawlike fashion occur together with a B which causes as R does not amount to the M causing the R.[28]

---

[24] p. 530.

[25] p. 531.

[26] p. 532.

[27] We ignore the complication arising from the possibility that "Cxy" is not primitive but definable by some other primitive which is non-Humean e.g. "x is a causal power to produce y".

[28] Even for a Humean, but there might be some way of obtaining suitable Humean causal relations out of the laws of the "Cxy"-less theory.

For example, let T_2 be: $(Ey)(Ex)(x=(U_i, t_i, L_i)$ & Ry & Cxy), and let T_1 be: $(Ey)(Ex)(x=(U_j, t_j, L_j)$ & Ry & Cxy), where (U_i, t_i, L_i) is a phenomenal event, and (U_j, t_j, L_j) its physical correlate. Then T_1 + $(x)(x=(U_i, t_i, L_i) \equiv x = (U_j, t_j, L_j))$ entails T_2, but T_1 + PSI does not.

This point is an ad hominem, but it does illustrate something: that IT and PSI differ as to consequences. There is no general guarantee that if T_1+IT entails T_2, then T_1+PSI does too. Let us be clear, though, what the ad hominem is. Brandt and Kim might try to reply that they have no need to consider a T_2 containing a non-Humean causal relation, because their claim is that IT and PSI have identical observational consequences, and only the "Humean component" of the causal relation can be observed. But if this is how IT and PSI are supposed to be equivalent, then it might not be that the sole reason for preferring IT to PSI is Ockham's Razor. There might be an argument for non-Humean causes, and if there is, then this argument will constitute a further reason for preferring IT to PSI. It is not simply the claim that IT and PSI have identical observational consequences that forces Humeanism on us, but this claim together with the insistence that the only reason for preferring IT to PSI is Ockham's Razor.

A Humean about causes might well be unmoved by this ad hominem. I proposed in the rest of this section to argue two things which perhaps will move him or her.

First, it is hard to see a good reason for thinking Ockham's Razor as here used to be "less than over-powering". Certainly ontological simplicity is a reason, other things being equal, for preferring one theory to another. Brandt and Kim do not argue that Ockham's

Razor is less than compelling. But the language they use suggest a reason they might have for believing it. Reasons for preferring the Identity Theory are contrasted with "ordinary scientific explanation"; the Identity Theory is "philosophical and speculative", and has a "false air of scientific respectability". Ockham's Razor used in support of the Identity Theory is "parsimony of a rather metaphysical sort".[29] The picture that emerges from these quotes is that of the distinctness of philosophy and metaphysics from science. Science is good, and metaphysics is less good. Ontological simplicity in the service of science is acceptable, but less so in the service of metaphysics.

I do not think the picture has much to recommend it. Even if we could separate out some issues which would clearly count as philosophical and not scientific, (perhaps a concern with conceptual questions?), there will remain a large middle ground between the two poles. Identity theorists in particular have spent a great deal of philosophical time emphasising the strong continuity between scientific questions and philosophical ones. The area of the Identity Theory, particularly, would seem to be one falling within this middle ground. Of course, philosophers have concerned themselves with conceptual questions when discussing the mind-body problem, though not only conceptual questions; but then to suggest that science is not or ought not to be partly a conceptual activity flies in the face of the facts.

Reject, then, the idea that Ockham's Razor in the service of IT is somehow not quite as strong as it

---

[29] All these quotes are from p. 534.

can be. Ockham's Razor is not however the sole reason for preferring IT to its PSI-based alternative, contrary to what Brandt and Kim claim. PSI together with the denial of psycho-physical identity is committed to emergent laws, and on at least two different accounts of it, the Identity Theory is not. And that is a reason for opting for IT.

Smart claimed as much in "Sensations and Brain Processes" when he said that the difference between the Identity Theory and Epiphenomenalism was that the latter required "a large number of irreducible psychophysical laws ... of a queer sort, that have to be taken on trust" or in other words, "nomological danglers". Brandt and Kim concede that Epiphenomenalism and the PSI without IT are committed to such (emergent) laws, and also concede that if this is a difference from IT, then there is "no question whether a rational person must accept it".[30] They seem to think that IT is likewise committed to emergence. The passage that most clearly seems to give their reason for thinking this (it gives their reason for rejecting the above point of Smart's) is as follows:

> there is no diminution of laws on IT; each particular psychophysical identity, in our view, logically entails a correlation law, and in this sense the identity is at least as queer as the correlation law.[31]

This contains several mistakes. One is to think that the problem about emergence has something to do with the number of laws rather than, as we have explained it, the deducibility of laws about things from laws about

---

[30] p. 533.

[31] p. 533.

their parts (or in Smart's words, their "irreducibility"). Another is to think that if p is queer to a certain degree, and q entails p, then q is queer to at least that degree. One cannot imagine what calculus of queerness Brandt and Kim have in mind. It is just not true that if something is queer, and we subsequently discover that something else's being true is responsible for the first thing's being true, then the first thing remains queer and the second takes on its queerness as well. It is one way that we reduce our wonder about the world, to find things like the second that are not so queer and then revise our ideas about the first. Perhaps the main mistake, however, is one that they make in their previous articles as well. That is to compare IT with PSI. PSI by itself is incomplete. Providing that it does not eliminate Ms, the final theory to be complete should tell us whether Ms are identical with Bs, or not. PSI is neutral on this fact. So the theories to be compared are IT and PSI + ~IT. In their earlier articles they thought that there were some advantages in remaining neutral about the truth or falsity of IT. But a theory which is neutral on this point cannot hope to be the final theory.

It is clear, I think, that theories like Epiphenomenalism and Parallelism are committed to emergent laws. The question is whether the Identity Theory is. One would certainly think prima facie that it is not. Here is a place where the machinery we introduced in the earlier chapters can help us.

Suppose that a topic neutral analysis has been successful. (Brandt and Kim do not consider the strategy of the topic neutral analysis.) Mental predicates are predicates of a number of candidate reducing theories.

Whether or not the Identity Theory requires emergent laws, then, is a matter of whether any of the candidates which are also Identity theories, do not have emergent laws. It is easy to imagine that they do not. As things stand at the moment, there appears to be no evidence from neurophysiology that the brain needs emergent laws as part of its description (although, c.f. Part Two). Far from its being a necessary consequence of the Identity Theory, then, that there are emergent psychophysical laws is unlikely. Brandt's and Kim's position only appears plausible if you think of the mental predicates as not being analysable into neurophysiological terms; for then the biconditionals look more as if they have come out of the air, cannot be derived from neurophysiology and so must be emergent. But it is the strategy of the topic neutral analysis to show that mental predicates are predicates which can belong to neurophysiological theories, so that the laws, admittedly contingent, relating mental states and brain states can be derived from it.

Even if a topic neutral analysis is unsuccessful, still all is not lost. For the topic neutral analysis is the only avenue for avoiding the prima facie problem of mental predicates that we have explored so far, and there are other avenues. An avenue we have not explored is whether we might not contingently identify the properties expressed by mental predicates with physical properties. If this were possible, then there would be on the face of it a good reason for holding mental predicates to be eliminable from the final theory, namely that all the facts about what exists and what properties they have can be expressed without those predicates. Let us suppose that this elimination can be made, and let us suppose

further what seems reasonable, that such eliminability is consistent with IT. (The further assumption, in this context, would seem to be consistent with what is expressed by IT.)

What is important for our purposes is eliminability. The further assumption that eliminability is consistent with IT is made so that the argument can be seen to bear directly on Brandt and Kim. But if the final theory contains no mental predicates, then, other things being equal, there is no reason to suppose that there are any emergent laws in nature! The grounds we might have for supposing that there are, would be the existence in our theories of troublesome mental predicates. But if the predicates are eliminable, there are no laws concerning them in the final theory to trouble us.

I conclude then that Brandt's and Kim's attempt to weaken the value of Ockham's Razor is unsuccessful. Ockham's Razor is something which is necessary to make belief in physicalism and IT reasonable, but it is no weak principle. Correctly applied, and in conjunction with other Principles of Method like the above (namely, that we should avoid emergence), it does make belief in IT reasonable.

The discussion has brought out a couple of possible alternative strategies for the physicalist, namely to eliminate mental predicates, or somehow to identify mental properties contingently, i.e. without analysis, with physicalistically acceptable ones. We will be returning to discuss those strategies later in Part One. In the next chapter, however, we digress in order to explore a strategy that many physicalists seem to have thought

would solve their problems for them, adverbialisation. We now turn to this.

CHAPTER FOUR. ADVERBS

1. Introductory.

So far, we have done several things. In Chapter One, we set up a definition of physicalism using the idea of the set P of physicalistically acceptable predicates, and we inquired about the membership of P. In Chapter Two, we gave definitions of reduction and elimination by means of biconditionals and looked at one strategy for dealing with problematic mental predicates, the topic neutral analysis. We examined attempts at this sort of analysis in the style of Smart's original version, and rejected them. In Chapter Three, we looked at an improved version of this sort of analysis which uses the idea of Ramsey Sentences. We concluded that it was not so easy to see whether common sense psychology was consistent with physicalism, but indicated that there might be a dualist argument which circumvented answering this question. It was also suggested that there might be other ways for the physicalist to avoid dualism, but discussion of those ways was deferred until later chapters. In the rest of Chapter Three, a couple of problems connected with Ockham's Razor were discussed, and it was concluded that at the present time physicalism is unable to dispense with Ockhamist arguments, but that, correctly applied, arguments relying on these and similar Principles of Method are powerful ones.

Now someone might wonder why Smart and others seem to have thought the topic neutral analysis necessary. Take a predicate like "x has a green afterimage". Surely, it might be said, the trouble for materialism

arises from the alleged facts that this predicate is sometimes true of some things, that it appears to commit one to the existence of some green things, afterimages, and that there are no green things in the brain. Smart has employed the topic neutral analysis in an attempt to show that the second alleged fact is only alleged, and not a fact. But we could, it might be suggested, achieve the same result as Smart just by denying that the predicate has the structure it seems to have in virtue of which its use commits us to the existence of green things.

How would one go about denying that "x has a green afterimage" commits us to the existence of afterimages which are green, without performing some sort of analysis of the predicate? In this chapter, we will be looking at a claim that a number of philosophers have made, that one can achieve this result by employing the device of adverbialisation.

One place to start is a passage from an article by Brian Medlin. Medlin in this passage was speaking about the predicate "x has a sense impression" rather than the predicate "x has an afterimage", but there is no doubt that his strategy if successful would be sufficient to dispose not only of sense impressions and afterimages, but also of pains, tingles, itches and their troublesome properties. He wrote

> It is true that if I say that there are sense impressions and that sense impressions may be blue and continuous, then I am in trouble. But it is one thing to say that I have sense impressions: that can be regarded as a philosopher's colloquialism. It is quite another thing to say that there are sense impressions which I have. When I say that I have a sense impression of my tobacco packet, the expression "have a sense expression of a tobacco packet" may be taken

> as a non-relational predicate, as though all the words in it were hypenated together. There is no need to construe my remark as of the form
>
> (Ex)(x is a sense impression of a tobacco packet and I have x)
>
> If we do not suppose that there are any sense impressions then there is nothing to possess the phenomenological properties of blueness and continuity.[1]

It is not absolutely clear what Medlin was trying to say in this passage. One very reasonable interpretation of the words "non-relational" and "hypenated" is that the predicate in question is to be regarded as semantically unstructured. That is to say, according to this interpretation Medlin is claiming that the meaning of the predicate is not (even partly) determined by the meanings of any parts of it. To put it another way, he is claiming the predicate is primitive or unanalysable. On the other hand, later in the same article he proposes that a person's using the predicate "means amongst other things, that he is in a condition which (typically) arises when he is looking at something blue and 'continuous' and which (typically) gives rise to the belief that there is before him something blue and continuous."[2]

Nothing prevents a person from contradicting themselves. An application of the principle of charity however might lead us to think that perhaps Medlin was using "means" in the sense in which we say "The presence of clouds means rain"; that he intended not that the predicate means whatever, but (perhaps) that from the person's using it we can reasonably conclude that ...

---

[1] Medlin 1967 p.107.

[2] p.108.

Whatever we make of this, the first suggestion is certainly interesting in its own right, and so I will suppose that that is what Medlin meant.

This is Interesting, but a very little reflection shows it to be inadequate as a way of avoiding quantification over mental objects. The suggestion that, say, "green" in "x has a green afterimage", or (Frank Jackson's example)[3] "in his leg" in "x has an ache in his leg" contribute nothing to the meanings of the respective predicates, is unbelievable. There is clearly a meaning relation between "x has an ache in his leg" and "x has an itch in his leg", and it is clear also that part of what is responsible for that meaning relation is the occurrence of "in his leg" in both. It follows from this that the occurrence of "in his leg", (with the meaning that it has, of course) in "x has an ache in his leg" is relevant to the meaning of "x has an ache in his leg", and that therefore an adequate semantic analysis of the predicate would bring this out. This, needless to say, would be impossible if "x has an ache in his leg" were unanalysable i.e. were semantically structureless. The same evidently goes for "x has a green afterimage".

While Medlin's suggestion is too crude as it stands, there are better proposals which derive, though not always by intention on the part of their authors, from as attempt to give a semantic structure to the predicates in question without incurring troublesome ontological commitments. In order to understand them, let us digress briefly into some recent work by Donald Davidson on the nature of action sentences.

---

[3] Jackson 1974 p. 9.

2. Davidson on Actions.

An interesting method for investigating whether the use of our language commits us to the existence of entities of a certain sort has recently been developed by Donald Davidson. In "The Logical Form of Action Sentences"[4], he argued that actions exist, on the grounds that the entailment relations that hold between sentences asserting that certain actions were performed, can be accounted for only by regarding those sentences as quantifying over actions. Consider the sentence "Hamlet killed Polonius with a knife". If we treat this as involving the three place predicate "Kxyz" for "x killed y with z", and so translate it as "(Ex)(Nx & Khpx)", then it becomes entirely mysterious why "Hamlet killed Polonius with a knife" should entail "Hamlet killed Polonius". As a solution, Davidson proposed to treat "killed" in "Hamlet killed Polonius" not as a two place predicate, but a three place predicate, where the variable in the extra places takes as its values actions. "Hamlet killed Polonius" translates as "(Ex)(Killed(h, p, x))", and "Hamlet killed Polonius with a knife" becomes "(Ex)(Ey)(Killed(h, p, x) & Knife(y) & With(x, y))"[5], from

---

[4] Davidson 1967

[5] Davidson's original translation gives rise to certain difficulties which can be avoided while remaining within its spirit. When a sentence has a place that can be bound by a quantifier, we are able to replace singular terms occurring in that place by "something". Thus "Hamlet killed Polonius with a knife" entails "Something killed Polonius with a knife", "Hamlet killed something with a knife", and "Hamlet killed Polonius with something". Equally, however, it entails "Hamlet did something to Polonius with a knife", and yet Davidson's construal, "(Ex)(Ey)(Killed(h, p, x)&N(y)&W(x, y))" makes the entailment to (indeed the construal of) this latter sentence mysterious. A way out is to translate it as "(Ex)(Ey)(Killing(x)&By(h, x)

which it is obvious that the expected entailment holds. If we do treat action sentences this way, then we commit ourselves to the existence of actions; and since there seems no other way to deal with the entailments save by ad hoc fiat, it seems reasonable to conclude that actions exist.

Part of the interest of Davidson's method lies in its broad applicability. By this I do not mean that the style of argument is always or even ever conclusive. Clearly, though, if we are looking for the ontological commitments of a class of sentences we should be looking at the meaning of members of the class, and entailment relations will be a good place to start to get a clue as to meaning. So let us see what we can extract from the entailment relations that <u>afterimage</u> sentences bear to one another.

There are two types of entailment I wish to focus on. Type I entailments are exemplified by such arguments as

x has a green square afterimage

$\therefore$ x has a green afterimage

or in general, the entailments from "x has an $F_1 \ldots F_n$ afterimage", to "x has a $G_1 \ldots G_m$ afterimage", where in place of the "F_1", "F_2", etc., go words like "green", "square", and the $G_1 \ldots G_m$ all occur in the list $F_1, \ldots, F_n$ and in the same order.[6] Type II

& Of (x, p) & With (x, y))". This also gives an account of the entailment to "Polonius was killed", and various other consequences in the passive voice.

[6] We also include the case where $\{G_1, \ldots, G_m\} = \Lambda$ i.e. entailments like the one from "x has a green afterimage" to "x has an afterimage".

entailments are exemplified by the argument

x has a green square afterimage

∴ x has a square green afterimage

i.e., where we can re-arrange the "F_1", ..., "F_n" to get the conclusion.[7]

Davidson's method appears to bring us immediate success with Type I and II entailments. If we translate "x has a green square afterimage" in such a way as to quantify over afterimages, arguments of Type I and II are easily validated by standard rules in first order functional calculus. Our first example becomes

(Ey)(Ay & Gy & Sy & Hxy)

∴ (Ey)(Ay & Gy & Hxy)

and our second example becomes

(Ey)(Ay & Gy & Sy & Hxy)

∴ (Ey)(Ay & Sy & Gy & Hxy)

On the other hand, construing (as Medlin seems to do) "x has a green afterimage", "x has a green square afterimage", and "x has a square green afterimage" respectively as "Fx", "Gx", "Hx", with no internal structure, makes the above two entailments mysterious.

When I say that it makes Types I and II mysterious, I do not wish to imply that deductions and entailments always need an explanation. Nonetheless, merely <u>stipulating</u> that "x has a green square afterimage"

---

[7] We will speak rather indiscriminately of valid arguments and entailments. We presuppose that the argument <u>p</u> is valid iff p entails c.

∴ c

entails "x has a green afterimage", has a very ad hoc ring to it. Jackson began to put his finger on this feeling of ad-hocness when he pointed out that meaning relations between some of the sentences in question seem to be due to a semantic component associated with certain of the words, and not others, in the sentence. Here we have cases where the meaning relations in question are rather more well-known than those of Jackson's example, namely entailment relations. There is, for example, a large class of entailments (type I entailments) which seem to be entailments for the same sort of reason as the actual example we gave is e.g. "x has a red square afterimage" entails "x has a red afterimage". But treating "x has a green square afterimage" as semantically structureless, and so also on parity "x has a red square afterimage", means that we cannot give any account of this sameness of structure between the two entailments: we should have to be stipulating independently that the latter was an entailment. Another way of making this point is to notice that "x has a red square afterimage" differs from "x has a green square afterimage" in replacing "red" for "green". But this replacement has certain effects on the entailment relations and not others: when the replacement is made, the new sentence does not entail "x has a green afterimage" and the old one did. On the other hand, the new sentence does entail "x has a square afterimage", while the old one did. It seems then that this systematic variation in semantic properties of the predicates is at least partly due to a semantic component associated with the place in the predicate that "green" and "red", respectively, fill.

Unless, then, we can find some account of the structure of the predicates which is different from an account which quantifies over afterimages, we appear to be committed to the existence of such entities. And that, as we have pointed out, constitues a prima facie danger to materialism. We will be returning to consider this point, and with it a further point about treating the predicates as structureless, later. Now, however, we will turn to the attempt to give the predicates a structure different from the one suggested above.

3. Adverbial Constructions and Mental States.

Quite a few authors writing on epistemology and the mind-body problem seem to have thought that objectionable mental objects like afterimages can be avoided by some kind of adverbial construction. Two things were usually common to the accounts. First, instead of saying e.g. that in perception a person sometimes sensed a red sense-datum, they would say that a person sensed redly (that is, the accounts would employ a word with the syntactic appearance of an adverb). Second, the accounts typically claimed that the language being analysed did not involve the commitment to any objectionable entities, for example sense data.

The following philosophers, among others, have proposed accounts which they have called "adverbial" accounts, in order to facilitate the denial of mental entities of some sort: C.J. Ducasse, R.J. Hirst, Roderick Chisholm, James W. Cornman.[8]

---

[8]Ducasse 1951, Hirst 1959, Chisholm 1966, Cornman 1971. Hirst's theory is complicated by the fact

These accounts differ in one key respect from the view that we attributed to Medlin. They differ in that they (implicitly at least) recognise that the predicates in question have a semantic structure.

They recognise that the predicates have a structure, because they employ syntactic forms which are adverb-like. "x is sensing redly" is different in meaning from "x is sensing greenly", and it is clear from the form of words that this is intended to be due to some semantic difference between "redly" and "greenly". This is evidently an improvement on Medlin's suggestion.

However, as we will see when we look at Chisholm's reply to Davidson's argument, the recognition that such predicates have a structure goes no way at all to showing what that structure is. Moreover, the bare syntactic forms "x is sensing redly" or "x is afterimaging redly" give no clue as to why there should be systematic entailment relations between the predicates, and what they should be.

There is another way of showing that an account of the semantic structure of the predicates is necessary. For let us observe that locutions like "x is sensing greenly", "x is afterimaging greenly and squarely" are technical. They did not occur in the English language before they were introduced by the philosophers who wished to use them. Being technical, we cannot be expected to comprehend them straight off: we should like

---

that he holds that mental states have an inner and outer "aspect", and it is only in their outer aspect that mental states are adverbial. We will not discuss this.

an account of their meaning. And an account of their meaning is ultimately an account of their semantic structure.

These points can be briefly summarised as follows. Predicates like "x has a green afterimage" have meaning relations to one another, particularly entailment relations. The entailment relations do not appear to be "brute force" entailments, but hold in virtue of a semantic structure of the predicates. Authors who have employed an adverbial construction of the predicates have to a man failed to give an account of the semantic structure, thereby rendering the entailments mysterious. This failure is even more serious, for we cannot easily understand someone who recommends that we utter the form of words "x is afterimaging greenly" in order to solve a philosophical problem. We should like to know the meaning of what he is saying.[9]

Not every philosopher who has employed adverbial constructions for <u>some</u> purpose has failed to see that typically such constructions require attention to their meaning. Chisholm, in a reply to the Davidson argument about events, attempts to use adverbial constructions and has something to say about their meaning.[10] We will look briefly at what Chisholm says,

---

[9]For example, it <u>might</u> turn out that "x is afterimaging greenly" is only another form of words for "(Ey)(y is an afterimage & y is green & x has y)". Just changing the syntactic form is useless.

[10]As far as I can determine, Chisholm has not made the connection between what he saw was necessary for adverbial accounts of action sentences, and his own well-known adverbial account of sentences apparently describing perceptual sensations.

for it will serve to lead into a more thorough-going account of the semantic function of adverbs. That account, in turn, will enable us to put the adverbialist's position more strongly.

4. The Semantics of Adverbs.

In "States of Affairs Again",[11] Chisholm argues against Davidson that an alternative account of action sentences can be given which preserves their entailment relations, does not involve quantification over events or actions, and hence has the advantage of being ontologically neutral. Chisholm admits that we do use sentences which appear to quantify over events, but he dismisses this more or less as a colloquialism.[12] The problem, which Davidson sought to solve by quantification over events, was to account for the entailment from (2): "Sebastian strolls in Bologna at 2am" to each of (3): "Sebastian strolls in Bologna", (4): "Sebastian strolls at 2am", and (5): "Sebastian strolls". Chisholm asks: "In virtue of what principles may we say that (2) above, entails (3) and (4), and that (3) and (4) entail (5)?"[13] Chisholm answers his own question by distinguishing between "genuine adverbial expressions" and "pseudo adverbial expressions". The distinction is made by examples - of the former, "swiftly", "in Bologna", "at 2am"; of the latter, "potentially", "apparently", "in his dreams", "in the imagination".

---

[11] Chisholm 1971. See also Chisholm 1970.

[12] Chisholm 1971 p.181, 3rd. para. Davidson was not, of course, arguing for events on the basis that we can say "There occurs that event which is the strolling of Sebastian in Bologna at 2am."

[13] Ibid., p.181.

Genuine adverbial expressions[14] are said to have the following effects on sentences:

> Consider any predicative or relational expression E (e.g. "strolls", "is red", "is larger than"); then consider any well-formed sentence S obtained just by adding to any such E either terms, or quantifiers and variables, or both. Then (i) the result of prefixing E to any adverbial expression, or to any conjunction of any number of adverbial expressions, will be a well-formed sentence S^1, (ii) S^1 will imply S, (iii) S^1 will imply the result of prefixing S to any adverbial expression that appears in S^1 or to any conjunction of adverbial expressions that appear in S^1, (iv) the result of prefixing S to any disjunction of any number of adverbial expressions will be a well-formed sentence S^2, (v) S^2 will imply S, and (vi) the result of prefixing S to any adverbial expression that appears in S^2 or to any disjunction of adverbial expressions that appear in S^2 will imply S^2. [15]

Having claimed that genuine adverbial expressions have these effects on entailment relations, and that there are some genuine adverbial expressions, Chisholm then says that

> ... (3) entails (5) in virtue of (ii), and that (4) entails (5) in virtue of (ii). And, assuming that in (2) a conjunction sign is left tacit between the two adverbial expressions 'in Bologna' and 'at 2am', we may say that (2) entails (3) in virtue of (iii), and that (2) entails (4) in virtue of (iii).[16] (my emphasis)

---

[14] In the quote, Chisholm calls them "adverbial expressions". I think this is significant in that it suggests that Chisholm thinks that pseudo adverbial expressions are not really adverbial expressions at all. And this in turn suggests that Chisholm thinks that the entailment of (ii) above is somehow natural and so does not need any further analysis.

[15] Ibid., p.182.

[16] Ibid., p.182.

This is intended by Chisholm to be an alternative to Davidson's explanation of the entailments. One could imagine someone similarly holding that our Type I and Type II entailments between mental predicates are explicable by the same sort of device: Chisholm is speaking about adverbial constructions already existing in English, but there is no reason why something which is in fact adverbial should not have its adverbial nature made more manifest by employing a syntactically adverbial construction whose semantic properties are the same as (i) - (vi) above. To put it another way: following Chisholm, someone might claim that "greenly" and "squarely" in "x is afterimaging greenly and squarely" are genuine adverbial expressions, and that <u>therefore</u> "x is afterimaging greenly and squarely" has all its correct Type I and Type II entailments, with no problem about quantification over unwanted afterimages.

This would be an improvement over the earlier accounts, but it is still not yet good enough. The first thing to see is that this way could be taken with any <u>alleged</u> entailment, even if it were not an entailment. For example, it is important to notice that adverbial expressions (and by that I mean expressions with the syntactic form or grammatical positioning that we loosely call adverbial) do not always have the property of making <u>p, adverbially</u> valid. Chisholm's examples of $\therefore p$ "pseudo-adverbial expressions" suffice to make this point: "potentially", "apparently", "in his dreams", "allegedly", even "possibly". Chisholm seems to have thought that this makes them not <u>really</u> adverbial, instead of drawing the correct conclusion that adverbial expressions have a variety of entailment properties. (Even if we agreed with Chisholm on this point, we should not find any

comfort in believing that "greenly" in "x afterimages greenly" has the right entailments, because we would have to establish the further point that "greenly" is a genuine adverb, its syntactic appearance being no guarantee.)

But once we reach this conclusion, we can also see that it does no good to say that the entailment holds "in virtue of (ii)". All (ii) says is that the entailments from a certain class hold. We could equally say that the entailments from a suitable class held even if one of its members was not an entailment at all. We could cook up a rule like (ii) to "explain" any alleged entailment, but it would not be an explanation without some reason to think it true. The only reason Chisholm seems to have for thinking (ii) true of the adverbial expressions he is interested in, is that it holds of genuine adverbial expressions by logical necessity. So, then, of any one of these problematic expressions we should like to know what reason there is to think it genuine and not pseudo. We could not say "because the relevant entailments hold" because that would be going around in a circle. Similarly with the adverbial removal of afterimages. It would be no good telling us that the adverb "greenly" falls in a certain class of adverbs determined by their entailment properties in order to justify the entailments that the adverb has. We need some way of breaking into this circle, or the entailment relations in question are just as mysterious as they would be under Medlin's suggestion.

I do not insist that all entailments will need justification or explanation.[17] However, the entailments

---

[17] One way in which the problem about synthetic

we are discussing seem to arise from semantic structures and an account of semantic structures is what will justify entailment relations insofar as they need justification. For an account of semantic structures will be at least an account of the truth conditions of those sentences. It will be an account of how the truth of complex sentences arises from the semantic properties (truth, satisfiability, etc.) of their parts. With an account of the truth conditions of sentences like "Sebastian strolled in Bologna at 2am" we will be able to explain its entailment properties, for entailment relations are those relations which invariably lead from truths to truths.

There is thus a crucial difference between what Davidson claims about the structure of action sentences and what Chisholm claims. Davidson offers an account of the truth conditions of action sentences. (He does this by translating them into first order functional calculus, the truth conditions for sentences of which are well-known.) Chisholm contends that quantification over objectionable entities is unnecessary to preserve the entailment properties in question. Because he does not provide an account of the truth conditions of Davidson's problematic sentences, however, his claim is empty. It amounts to the mere assertion that the entailment relations could hold without the sentences having the semantic structure that Davidson says they have.

The same point holds for any adverbial account

---

a priori propositions arises is that some entailments seem to need an explanation but none of a certain sort seems available. e.g. the entailment from "x is red" to "x is coloured".

of mental talk which seeks to do away with afterimages and the like. Afterimage sentences have entailment relations. It is useless to just assert that these relations could hold without quantification over after-images. We should like to know how this could be and what part semantically relevant words like "red", "square" play in contributing to these relations.

All is not lost for the adverbialist, however. In fact, not much is lost at all. A great deal of discussion of the semantic properties of adverbs has taken place in the last few years.[18] The most recent and the most thorough-going investigation has been made by Malcolm Rennie.[19] We will now look very briefly at it, in order to see how it might be used to help the adverbialist about mental objects.

Adverbial expressions modify more verbs than just action verbs. If, with Davidson, we are going to take the entailment relations between sentences with action verbs in them as sufficient reason for having actions in an ontology, then we should presumably also have to include things like states, on the grounds that the inference

Hamlet was ill from melancholy

$\therefore$ Hamlet was ill

can be justified using the machinery of first order logic only by quantification over the state, illness. But then, as Romane Clark puts it, "the finger is out of the dike".[20]

---

[18] See e.g. the bibliography to Rennie 1974.

[19] Rennie 1974.

[20] Clark 1970 p.311.

We do not merely talk about what Jones does, but also what he is and has, "the states, offices and natures of things".[21] There is strong motivation, therefore, for looking for some sort of extension of predicate calculus which will make adverbial inferences come out valid and yet give an account of the semantic structure of adverbial sentences which avoids the ontic avalanche.

It is easy enough to provide a formalism which syntactically distinguishes adverbial constructions from ordinary predicates. We could lay down a stock of n-adic adverbial expressions[22], a typical one might be "f^i_j ( )(x_1, ..., x_i)" (the superscript for the adicity, the subscript for the jth. expression of that adicity). Then if we take the expressions of the first order functional calculus and call them wffs, we can specify that if α is a wff, f^i_j (α)(x_1, ..., x_i) is a wff. Syntactically adverbial constructions would consist in nestings around a central "core" expression of ordinary predicate logic. This is a simplification of Rennie's system, which employs type-theoretic indexing of expressions, but we neglect this complication as unnecessary for our purposes.

What about the semantics of such expressions? Adverbs modify things, and semantically we take the word "modify" literally i.e. to mean "change". Adverbs take expressions with certain truth conditions or satisfiability conditions and change those conditions. Take for example the class of adverbs which we can call

---

[21] Clark, Ibid., p.311.

[22] We should need adverbs of adicity greater than 0 so as to treat the entailment relations of prepositional phrases e.g. "with a knife" in "Hamlet slew Polonius with a knife".

predicate modifiers because they semantically change predicates.[23] Now a predicate modifier can modify a variety of predicates, even of varying adicity, without changing its meaning. One can run, jump, stand still, and kill Polonius, all slowly. In standard first order semantics, predicates are associated with their extensions. So adverbs for Rennie are associated with functions, which take extensions to extensions e.g. "slowly" will be associated with a function which takes, among others, the extension of "x runs" (a class of individuals) to another extension (the class of individuals that run slowly), and which takes the extension of "x jumps" to the extension of "x jumps slowly". Because adverbs are associated with functions like this, they are able to be seen as single semantic units which can operate on varieties of semantic material to produce varieties of results (changes, modifications). Truth of a sentence with no free variables is then defined in terms similar to the way it is defined in ordinary predicate logic. For example, a modified predicate is treated semantically as another predicate i.e. something with an extension which is the result of modifying the original extension of the unmodified predicate.

This regretably sketchy and oversimplified account will perhaps be aided by some examples.[24]

---

[23] Of special relevance to us because we are looking at how "greenly" can modify "x afterimages".

[24] One way that it is simplified is by ignoring the fact that Rennie has intensionalised his semantics by introducing possible worlds, evidently necessary if only to be able to treat "possibly" as an adverb.

Example 1: To validate "$M(\phi x) \rightarrow \phi x$" [25]

Let us be clear what we mean by "validate" here. For some adverbs, the inference $\frac{\alpha\text{, adverbially}}{\therefore \alpha}$ is valid (as we have been using the term), for others not. The validity of the inference, therefore, depends on the semantic properties of the particular adverb. (It will also depend on the semantic features of α, if only because α might itself contain adverbs, but we will not concern ourselves with this complication.) This being so, the semantic properties of adverbs cannot be investigated by treating adverbs in the way in which predicates are treated in first order logic, as schema or as allowing uniform substitution. If we did this, then a semantic condition necessary to validate the above would make the inference valid for all adverbs.[26] First order logic is intended to investigate the properties of terms qua predicates or relations, and that is why substitution is permitted in theorems. When it comes to the particular properties (logical or otherwise) of particular predicates, they must either be treated like the connectives, as something like logical constants[27], or in a theory, i.e. something not a logic at all. A useful example of the latter is the "logic" of relations. The argument schema $\frac{Rxy}{\therefore Ryx}$ is not valid in first order logic, but there are relations of which it is a matter of their

---

[25] For the examples we ignore sub- and superscripts.

[26] This point was apparently missed in Rennie and Malinas 1970, and Rennie 1971, in which a variety of systems are proposed for validating all inferences, even unwanted ones, of a certain form.

[27] An example where the first is done for a class of adverb-like words, is modal logic.

meaning that such arguments are valid e.g.
x is the same thing as y / ∴ y is the same thing as x . What is often done is that a theory is constructed, with a language containing only relational constants e.g. "Rxy", and axioms are laid down for them e.g. "(x,y)(Rxy → Ryx)", and what can be proved from these axioms using as "background" first order logic and what semantical structures are generated, is looked at. Within this approach, no distinction can be made between axioms which have the status of logical truths about the predicates being investigated, and axioms which are just being proposed as true. The theory of partial orderings does not look any different in principle from axiomatised Mendelian genetics. Nevertheless, we can regard our investigation as an investigation of the logical properties of symmetric relations just because we do have in mind that the semantic properties we uncover will be regarded as an account of the truth conditions of sentences with the relation in them as a matter of the meaning of the relation.

It is similar with adverbial modifier theory. Modifiers are treated as constants whose logical properties are to be looked at by applying semantic procedures which are not the same for all modifiers, but the semantics are to be regarded as part of the meaning of the modifier. In example 1, "ϕx" is intended to be a parameter: "M" is required to operate on any extension to give the desired entailment. In the semantics then, this amounts to requiring that the function associated with "M" operate on any extension we might choose for "ϕx".

If, for convenience, we let our universe of discourse be the set N of natural numbers, extensions

for "ϕx" will be subsets of N, so the function f associated with M will be a function which takes subsets of N to subsets of N, i.e. members of the power set of N to members of the power set of N. Thus $f \in \mathbb{I}\,N^{(\mathbb{I}\,N)}$. Now it can be shown[28] that the theory of the modifier "M" with just the one axiom "$(x)(M(\phi x) \rightarrow \phi x)$" is complete with respect to the appropriate class of structures in which "M" is associated with an f that obeys the condition $f(X) \subseteq X$ for all $X \in \mathbb{I}\,N$. So then, an <u>example</u> of a modifier for which "$(x)(M(\phi x) \rightarrow \phi x)$" is true no matter which "ϕx" we choose, is provided by any function $f \in \mathbb{I}N^{(\mathbb{I}\,N)}$ which satisfies $f(X) \subseteq X$. One such is $f(X) = \{y : y \in X \ \&\ y \text{ is even}\}$ i.e. the function which picks out just the even members of X.

<u>Example 2:</u> To find examples of modifiers "R", "S" which obey the axioms

(1) $(x)(R(\phi x) \rightarrow \phi x)$

(2) $(x)(S(\phi x) \rightarrow \phi x)$

(3) $(x)(R(S(\phi x)) \equiv S(R(\phi x)))$

(4) $(x)(R(R(\phi x)) \equiv R(\phi x))$

(5) $(x)(S(S(\phi x)) \equiv S(\phi x))$

and for which (6) $(x)((R(\phi x) \ \&\ S(\phi x)) \rightarrow R(S(\phi x)))$ does not hold.

We must find functions $f_1, f_2 \in \mathbb{I}N^{(\mathbb{I}\,N)}$ obeying

(1)' $f_1(X) \subseteq X$

(2)' $f_2(X) \subseteq X$

(3)' $f_1 f_2(X) = f_2 f_1(X)$

(4)' $f_1 f_1(X) = f_1(X)$

(5)' $f_2 f_2(X) = f_2(X)$

but NOT (6)' $f_1(X) \cap f_2(X) \subseteq f_1 f_2(X)$

---

[28] Rennie 1971.

Let $f_1(X) = f_2(X) = X$ unless $X = \{1, 2, 3, 4\}$ or $X = \{1, 3, 4\}$ or $X = \{1, 2, 4\}$, in which case,

$f_1(\{1, 2, 3, 4\}) = \{1, 2, 4\}$ $f_2(\{1, 2, 3, 4\}) = \{1, 3, 4\}$

$f_1(\{1, 3, 4\}) = \{4\}$ $f_2(\{1, 2, 4\}) = \{4\}$

$f_1(\{1, 2, 4\}) = \{1, 2, 4\}$ $f_2(\{1, 3, 4\}) = \{1, 3, 4\}$

It is not difficult to show that (1)' - (5)' hold. Therefore, if we assign f_1 to "R" and f_2 to "S", (1) - (5) are true no matter which "ϕx" we pick. As to (6)', we have $f_1(\{1, 2, 3, 4\}) = \{1, 2, 4\}$ $f_2(\{1, 2, 3, 4\}) = \{1, 3, 4\}$, so $f_1(\{1, 2, 3, 4\}) \cap f_2(\{1, 2, 3, 4\}) = \{1, 4\}$. But $f_1f_2(\{1, 2, 3, 4\}) = f_1(\{1, 3, 4\}) = \{4\}$, so $f_1 \cap f_2 \not\subseteq f_1f_2$. If, then, we let "$\phi x$" be true just in case $x \in \{1, 2, 3, 4\}$ (i.e. if we set $\{1, 2, 3, 4\}$ to be the extension of "ϕx") and we assign 1 to the variable "x", we have that "$R(\phi 1)$ & $S(\phi 1)$" is true, but "$R(S(\phi 1))$" false, thereby falsifying (6).

How can this be used to help the adverbialist about actions or afterimages? The first thing to note is that Rennie's theory gives a framework for the truth conditions of sentences with adverbs in them. If "M" is associated semantically with a function taking sets of numbers to sets of numbers, and "Fx" associated with a set of numbers, then "M(Fa)" is true just in case "a" names something which belongs to the extension of "M(Fx)", and "(Ex)M(Fx)" is true just in case something belongs to the extension of "M(Fx)". This being so, the criticism that the adverbialist cannot account for the entailment relations between afterimage sentences no longer works. The adverbialist's claim will be that adverbs like "greenly" function semantically like functions, taking extensions like those of "x afterimages" and "x afterimages squarely" to other extensions, and the modified predicates will be true of objects when those

objects belong to the modified extension. Furthermore, the adverbialist can claim that it is a matter of the meaning of "greenly" that it takes extensions to subsets of those extensions, as in example 1 above. From this it follows that "x afterimages greenly" entails "x afterimages".

Since the adverbialist can give this sort of account, it follows that a Davidson-style argument for the existence of afterimages does not work. (It also follows that a Davidson-style argument to the existence of actions does not work.) For, let us recall, Davidson-style arguments rely on a demand for an account of the truth conditions of sentences, and entailment relations between sentences, other than an account which implies that such sentences entail the existence of the entities in question.

It might now be thought that Rennie's theory carries the day for adverbialists about actions and afterimages. We are faced with alternative semantic frameworks for action sentences and for afterimage sentences. Should we not choose that which is ontologically more parsimonious, the adverbial theory? In the next section we will look at this argument in the light of the strategy that an adverbialist might be able to adopt.

5. Methodological Considerations.

Let us try to say clearly what someone might intend to use adverbial constructions for classes of mental predicates for. (1) A person might be primarily interested, as Chisholm was in The Theory of Knowledge[29],

---

[29] Chisholm 1966.

in making a point about perception; in trying to produce an adverbial construction as an alternative to having to accept the existence in perception of sense data. (2) Alternatively, a person might be intending to use adverbs not in the service of a theory of perception, but in the service of materialism. They might feel that the existence of mental objects of a certain sort would threaten materialism, and so wish to dispose of them. Closely related to this aim, but distinct from it, is (3) the intention merely to dispose of various sorts of mental entities. It is distinct because whether or not there are green afterimages is logically independent of dualism. Dualism can be property dualism without the existence of dualist objects. Conversely, the existence of green afterimages in a wholly grey brain does not entail dualism, for "green" as a predicate of afterimages might have an analysis which makes such "greenness" compatible with the greyness of the brain. "green" applied to afterimages might not mean what it looks like it means.

For each of these three aims, for each of these three kinds of adverbialist, there are two different uses to which adverbs might be intended to be put. A person might be proposing their adverbial constructions to be bearers of the truth. They might be proposing their theory as the truth. They might think (1.1) that the correct account of perception is an adverbial one, they might think with Cornman (2.1) that an adverbial materialism is true, or (3.1) they might think that there are no mental entities, only "adverbial" mental properties of persons. On the other hand, a person might have something weaker in mind, namely((1.2), (2.2), (3.2)); to show that there is an alternative way of construing

mental predicates which shows that their use does not commit one to objectionable entities.

We are deferring questions about perception for a later chapter, though much of what we say here is relevant to them, so we will not discuss (1.1) and (1.2). Let us look first at the person whose aim is to deny mental objects, as in (3.1). We have already seen that an adverbial account is useless without a semantics, an account of the meaning of the adverbialised predicates. It is also true, however, that such an adverbialist is doubly committed to giving a semantics. This is because if they propose their theory as true, it is useless to offer opponents adverbialised sentences without some indication of what they are claiming about the world. They must say what the world would be like if their sentences were true.

Even then the job is not done. It is no use at all to offer your theory complete with semantics, and think that is an end to it. If we are to believe it, we will need some reason for thinking that it is true. In addition to presenting your theory, that is, you must argue for it.

One such argument might come from Ockham's Razor. It is an argument which one feels has some strength when we are measuring an adverbial theory of actions against Davidson's theory. But even though an adverbialised account of afterimage sentences carries less ontological commitments than one with real afterimages in it, this does not mean that we should immediately opt for adverbialisation. Simplicity is one consideration among others, nothing more. If we could find a knock-down argument for afterimages, simplicity would have to

be set aside. If we could discover real afterimages in our consciousness, for example, then Ockham's Razor should have no power to move us in the opposite direction.

There might of course, be other arguments for an adverbial theory. One such might be this. After-images are dependent entities. Logically, an afterimage cannot exist without someone to have it. But if an afterimage were distinct from the person who has it, even if it were a part of the person, then it could exist when no-one has it. Therefore, afterimages are not entities over and above people. Therefore, afterimages are not entities at all. The answer to this argument is twofold. First, where does the claim that afterimages logically cannot exist without owners come from? I, for one, feel inclined to deny it. Even if it were true, however, it would still not show that afterimages were not entities distinct from their owners. It is well known by now that sons logically cannot exist without at some time having had a parent, and that this fact does not show that sons are not fit entities to take one place of a genuine relation, the "parent-son" relation.

The above remarks apply not only to adverbial-isation to deny mental entities (i.e. (3.1)), but also adverbialisation with the aim of supporting materialism (i.e. as in (2.1)). As we said before, the two aims ((2.1) and (3.1)) are distinct but connected. At least one way in which they are connected is that someone who succeeds in the attempt to justify denying afterimages and pains, does not have to worry about their properties. If there are no green afterimages, we need not worry about trying to find the green things in the head or analysing away "green". If there are no pains to be in

legs, then that my leg was amputated yesterday and still hurts today does not require ghostly pains or ghostly legs of us.

There is, however, one point which it is important to bring out. Even if you succeed in giving your adverbial semantics and showing your theory true, you have still not yet saved materialism. The reason, in a nutshell, is that showing that a predicates does not quantify over entities of a certain sort, does not show that the predicate is a member of P. The point can perhaps best be made in relation to Medlin's suggestion that we considered earlier. If "x has a green afterimage" has no semantic structure, if it is - as one might say - unanalysable, then for it to be in P it would have to be one of P's primitives. A consequence of Medlin's suggestion, then, is that mental predicates are not members of P and so are not physicalistically acceptable.

This point seems to have been missed by James Cornman[30]. He correctly sees that it is not enough to give his adverbial theory, and that he has to argue for it. (His argument consists in refuting arguments in favour of contrary views which do claim that there are mental objects. Such a strategy would seem to be adequate when conjoined with Ockham's Razor.) But then he seems to think that in getting rid of mental objects he has done enough for adverbial materialism, apparently not noticing that the predicates he is left with are very strange indeed.

Of the six aims an adverbialist might have that we outlined at the beginning of this section, only two are left (since, to repeat, we are ignoring specifically

---

[30]Cornman 1971.

perceptual questions i.e. (1.1) and (1.2)) They are, briefly, either (3.2) to say that the existence of an adequate semantics for an adverbial theory shows that we need not accept Davidsonian arguments to the existence of materialistically unacceptable objects, or (2.2) to say that the existence of the adverbial theory shows that even if "x has a green afterimage" is true of someone, this does not obviously commit us to the existence of afterimages. I think that the point against Davidsonian arguments is well-taken, although, as we said before, this does not mean that there might not be independent arguments for Davidson-type conclusions. There might even be arguments which turned on examining the alternative semantics and finding it wanting somehow, even though not wanting in respect of entailment relations. (2.2) is also correct, though it seems to me that the treck through adverbial semantics was a hard way to find it out. Is it not just the point that you cannot conclude from the fact that a word occupies a noun position in an English sentence, that there is an entity or entities which the word denotes?

Our conclusions about the usefulness of adverbial analyses in defence of physicalism, then, are these. If we wish to use them to show that arguments from the entailment relations of mental predicates to dualism are inadequate, or that arguments from the existence of certain terms apparently standing for mental objects to dualism are inadequate, then we will be successful. On the other hand, if we wish to use them to show that physicalism is true, then we will need to couple them with other considerations, specifically metaphysical ones. Now in a forthcoming article, Frank Jackson gives an ingenious argument for a stronger conclusion about

adverbial analyses than these[31]. Jackson argues positively against adverbial analyses of e.g. afterimage sentences, and for the existence of mental objects like afterimages and pains. In the next sections, we will look at his argument. I will try to argue that when fixed up, it is a plausible argument for mental objects.

6. Jackson on Adverbial Theories of the Mental.

Jackson discusses two types of theories, which he calls "adverbial" theories and "state" theories. We will confine our attention to theories of the first sort, although our remarks will apply to theories of the second sort insofar as they can be seen as being attempts at giving ontological stiffening to an adverbial theory. Jackson poses a problem against adverbial theories: how are we to analyse "x has a red square afterimage" in an adverbial way so as to avoid commitment to afterimages? He gives three alternatives (1) "x afterimages redly and squarely", (2) "x afterimages redly squarely", (3) "x afterimages red-squarely". Let us take them in turn.

Because Jackson does not operate in a framework of well-defined semantics for adverbs, it is not absolutely clear what he means by "x afterimages redly and squarely". He does however give a clue. He says that this analysis "has the advantage of explaining the entailment from 'I have a red, square afterimage' to 'I have a red afterimage' : for it will correspond to the entailment from 'I sense redly and squarely' to 'I sense redly' ." (pp. 16-17). This latter alleged

---

[31] Jackson 1974.

entailment however is not at all obviously an entailment. As was said before, the adverbial locution here is technical, and needs its semantics explained before we can agree to the entailment.

It would be an entailment if we understood "x afterimages redly and squarely" as a straight conjunction, "x afterimages redly and x afterimages squarely", for no matter what we make of the conjuncts, at least the behaviour of "and" is uncontentious here. It is fairly clear that this is what Jackson has in mind, for he goes on to argue that this interpretation obliterates the distinction between "x has a red square afterimage and x has a green round afterimage", and "x has a red round afterimage and x has a green square afterimage": both come out as "x afterimages redly and roundly and squarely and greenly". (This would not follow unless you were allowed to rearrange "redly and squarely and greenly and roundly" to get "redly and roundly and squarely and greenly", which of course would be allowed on the conjunctive interpretation.) A slightly simpler way of making the point is to say that it obliterates the distinction between having a red afterimage and a square afterimage, and having a red square afterimage. The point is of course that on the conjunctive interpretation, Jackson's argument is sound, so this interpretation is inadequate for the adverbialist's purposes.

7. Jackson's Second Interpretation of "x has a red square afterimage.

Jackson explains what he has in mind for the second interpretation (i.e. (2) above) with the example "He spoke impressively quickly". Here, according to

Jackson, "impressively" does not modify the verb but the adverb "quickly". Similarly, we might interpret the "squarely" in "He afterimages redly squarely" as modifying "redly" and not the verb.

If this were the adverbialist's interpretation, Jackson argues, we should have the problem of deciding whether this was the correct order of modification of the adverbs, or whether it should be that "squarely" modifies the verb and "redly" modifies "squarely". And even if we could decide this in a non-arbitrary way, we should have to admit that "redly" means differently in "x afterimages redly", where it modifies the verb, from "x afterimages squarely redly", where it modifies the adverb.

It might be that this interpretation is what Jackson has in mind, but just writing two adverbs one after the other does not force this interpretation on us. Other examples suggest something more like a Rennie-type interpretation, in which the first adverb is regarded as modifying the predicate, and the second as modifying the whole modified predicate e.g. "x paints quickly badly". On this interpretation, it is false that "redly" means differently in the two different contexts. The whole point of treating "redly" semantically as a function is so as to allow it to be a single semantic unit which has different effects on different predicates. If we identify its meaning with that single semantic unit, then the same meaning (unit) will be associated with "redly" in the various contexts. Nor do we have to decide what modifies what. "redly" modifies something different in "x afterimages redly" from "x afterimages squarely redly", and likewise "squarely". It is not as

though if "redly" were to modify "squarely" in some context, it would have to do so wherever they occur. What modifies what depends on the semantic structure of the sentences in question, and there is no reason why "redly" should not modify "x afterimages squarely" in "x afterimages squarely redly" and also "squarely" modify "x afterimages redly" in "x afterimages redly squarely".

Jackson does not say that one would have trouble explaining the various entailments on his second interpretation, but he does omit, as he does not for his first interpretation, to say that one would not. On our Rennie-type interpretation, however, one would not have this trouble, at least insofar as explaining entailments comes to what we said of it in previous sections. Example (2) in section 4 was an example constructed to show that such entailments as (1) $R(\phi x) \rightarrow \phi x$

(2) $S(\phi x) \rightarrow \phi x$

(3) $R(S(\phi x)) \equiv S(R(\phi x))$

(4) $R(R(\phi x)) \equiv R(\phi x)$

(5) $S(S(\phi x)) \equiv S(\phi x)$

can be made to hold. Read "x is afterimaging" for "ϕx", "redly" for "R" and "squarely" for "S", (and in (1) and (2) also read "x afterimages squarely" and "x afterimages redly" for "ϕx"). Then we have some of the entailment relations between afterimage predicates. It is also worth noting that these entailments can be made to come out without (as in interpretation (1)) incurring the penalty of being unable to distinguish between "x has an R and S afterimage" and "x has an R afterimage and x has an S afterimage". If they were indistinguishable, then on the adverbial theory the latter would entail the former, but example 2 is an example where "$R(\phi x)$ & $S(\phi x)$" does not

entail "R(S(ϕx))".

This objection can also be made against a later argument that Jackson gives against various adverbial theories, the "complement objection". According to this argument, if we say "x afterimages F-ly and x afterimages non-F-ly" we will have to conclude that x afterimages F-ly and non-F-ly; whereas while it might be possible for x to have a square afterimage and a non-square afterimage, it is impossible for x to have a square non-square afterimage. Example 2, however, shows that it is not generally legitimate to move from "$M_1(Fx)$ & $M_2(Fx)$" to "$M_1(M_2(Fx))$", and so presumably not always legitimate to move from "$M_1(Fx)$ & $\bar{M}_1(Fx)$" to "$M_1(\bar{M}_1(Fx))$" (always supposing we have a clear sense for "$\bar{M}_1(Fx)$" other than " ~ $M_1(Fx)$"). At least, Jackson does not provide a reason for thinking that the inference holds in the special case which is favourable for his objection.

8. Jackson's Third Interpretation of "x has a red square afterimage.

Let us now consider Jackson's third interpretation (i.e. (3) above), namely "x afterimages red-squarely". The terminology derives from Sellars[32] (who does not explain it) but it is, as Jackson notes, suggestive. It suggests not a fusing of the whole predicate "x has a red square afterimage" as with Medlin, but a fusing of the modifiers. "red-squarely" is intended, we might think, to be semantically structureless. This will have the effect of making the entailment from "x afterimages red-squarely" to "x afterimages redly" mysterious, as with Medlin's theory. Jackson argues against it differently.

---

[32] Sellars 1968.

He says that afterimaging red-squarely is a "special case" (p.20) of afterimaging redly, that the former has the latter as a "component" (p.20). Another way of putting the same point is to say that what in the world (i.e. a property) corresponds to "x afterimages red-squarely" has distinct and distinguishable elements, which are associated with the "red" and the "square". But if this is so, those elements ought to be reflected in the semantic structure of the predicate by splitting up the semantic contribution of "red" and "square".

This argument has considerable persuasive force, although I am not sure how to go about explaining "components" and "elements". The reason for my uncertainty is that the sort of thing which (intuitively) corresponds to predicates, is properties; and the idea that the property afterimaging redsquarely should have as a component the property of afterimaging redly is a hard one to make clear.

Let us press on with Jackson's argument. We will return to this point later and try to make it clearer. The language "red-squarely" suggests to Jackson that afterimaging red-squarely is being thought of as a "fundamental mode" or "basic mode" of afterimaging. (p.22, 23). These seem to me to be alternative ways of saying "structureless" but there is a problem here. It was not so bad to talk about a predicate's being semantically structured or structureless, but what is it for a property to be structureless? An answer that intuitively recommends itself is "simple", or "without components or elements" and that of course brings us back to components and elements. Let us try to make do with that for a while.

Jackson has three arguments against saying that afterimaging red-squarely is a basic mode of afterimaging. His first argument is very similar to the one just given. It is: . if afterimaging red-squarely is a basic mode, then afterimagingredly would not be a component, which it intuitively is. (That he should argue thus reinforces the interpretation of "basic mode" as "without components".) His second argument is that if afterimaging red-squarely is simple, and if, on parity, afterimaging red-roundly is simple, then we have no account of the fact that someone who afterimages red-squarely thereby has something in common with someone who afterimages red-roundly. They just have different, simple properties, that is all. Jackson's third argument is that for no $n \geqslant 1$ can afterimaging $F_1, \ldots, F_n$-ly (the property corresponding to "x has an $F_1, \ldots, F_n$ afterimage") be a fundamental mode because of the indefinite number of things that can be said of an afterimage. We must admit that an indefinite number of things can be said about an afterimage (red, square, fuzzy, superimposed on the bookcase, to the left of a green round afterimage, having patches of crimson and patches of ochre ...). But now if we say that afterimaging $F_1, \ldots, F_n$-ly is simple, what can we possibly say about afterimaging $F_1, \ldots, F_n, F_{n+1}$-ly? Would we not have to analyse it, as in interpretation (1), as afterimaging $F_1, \ldots, F_n$-ly and afterimaging F_{n+1}-ly, with the consequent troubles of (1)?

Notice this: that Jackson's arguments are directed against a certain linguistic theory - an adverbial analysis of certain predicates - but much of his discussion is about properties, about what in the world is associated with the predicates. That the discussion might well be like this follows from what we said earlier about the

strategy open to any adverbial theorist. We were able, in example 2, to construct adverbial modifiers which had properties very similar to those desired for an adverbial account. But adverbial theorists, if they propose their theory as true, must say not just what the appropriate entailment relations are, but also what the world is like, and it is on this point that Jackson is attacking them.

As we said before, at least some of Jackson's arguments have a persuasive force. I do not think that they can be relied on too heavily, however, for they exploit notions like that of a simple property, that of a component to a property etc. It might be possible to get these notions going, but even if we do so it is by no means certain that they will be any more congenial to someone who believes in mental objects than to someone who believes in mental properties, particularly the notion of a simple property. If the one must answer questions about whether some of their properties are simple, so must the other.

This is not to say that if someone holds that afterimaging red-squarely is simple, he or she is not open to objections utilising simplicity of properties. It is to say that we must tread warily if we want to wield these objections against any old adverbialist. It goes without saying that Jackson does not intend his objections to be so general.

9. An Argument Similar to Jackson's.

After this note of caution, we will throw caution to the winds and try to use objections like Jackson's against adverbialists in general. We found, in our discussion of Jackson's interpretation (2), that within

Rennie-style semantics it was possible to construct modifiers with many of the logical properties we should want for adverbialism about mental predicates. Now a Rennie-style semantics is in a sense a framework for various metaphysics, much the way first order logic is. It gives us sets, individuals and, if we want them, possible worlds. And, just as with first order logic, there is no compulsion to have only particulars as our individuals. (We might wish to give the familiar account of properties as extensions in all possible worlds, but I do not want to decide that question here.) Let us ask, then, what sort of metaphysics would an adverbial theorist, and in particular a Rennie-type adverbial theorist, about afterimages propose?

It seems clear that they should be proposing that afterimaging redly, afterimaging squarely, afterimaging redly and squarely etc., are properties of people[33]. Any such proposal should also intend that the properties be non-relational, at least to the extent that the possession of any one of them would not involve the existence of afterimages. Such an account would have the previously noted advantage of not requiring us to worry about the properties of afterimages.

If our adverbialist believes just this much, then it seems to me that he is vulnerable to an argument similar to Jackson's. To see this, let us introduce the ordering <u>being partly made up of</u> in the class of all properties[34]. I can only introduce it by examples and to

---

[33]This is meant to include states as properties, but to be more general.

[34]This ordering is at least a partial ordering, since it will not be the case that for any pair of properties, they are identical or one is partly made up of the other. It

that extent the argument, like Jackson's, is weak. But it seems to me that there is an intuitive, albeit cloudy, sense in which some properties, like e.g. the property of being red and square, are made up from others, in this instance the properties of being red and being square. We need not suppose that the properties that do the making up are simple, but in this sort of case we will say that the one is partly made up of each of the others, and further that the others are components of the one.

Now the property of afterimaging redly and squarely does seem to be partially made up of the property of afterimaging redly, and also of the property of afterimaging squarely. First, this seems intuitively the case. Second, afterimaging redly seems to be a component because afterimaging redly and squarely has something in common with afterimaging redly and roundly and something <u>different</u> in common with afterimaging greenly and squarely. Further, the two different things in common when taken together seem to make up, and exhaust, afterimaging redly and squarely. What is in common in the one case is afterimaging redly, in the other afterimaging squarely. By a parallel argument, any property of afterimaging F_1,...,F_n-ly has as components afterimaging F_1-ly,..., afterimaging F_n-ly.

The properties afterimaging redly and afterimaging squarely are distinct properties, and so we would like to know how to make sense of the difference between a person who is afterimaging redly and afterimaging squarely when it is two afterimages, one red and one square, that the person has, and a person who is afterimaging redly and after-

---

is not intended that for any property there is some other property such that one is partly made up of the other.

imaging squarely when it is one afterimage, a red square one, that the person has. I cannot think of any way of marking this distinction that has the faintest shred of plausibility, unless it is a way that allows that in the first case there are different bearers of a property associated with "redly" from a property associated with "squarely", and in the second case, the same bearer. Ex hypothesi, the difference cannot be marked by a difference in which of the properties of afterimaging redly, afterimaging squarely and afterimaging redly and squarely that the person has, for in both cases the person has all three. Nor can the difference be marked by a difference in the bearer of the properties of afterimaging redly and afterimaging squarely, for there is but one person in both cases and it is precisely where there is but one person that the trouble arises. (If we want to distinguish between two people, one afterimaging redly the other afterimaging squarely, and two people with only one of them afterimaging redly and afterimaging squarely, we can do so easily. Significantly, we do it by distinguishing between the bearers of the properties.)

I suggest, then, that the only solution is mental objects to bear properties associated with "redly" and "squarely". We even, I suggest, have names for these entities. We call them "afterimages", and the property of them that is associated with the adverbial locution "redly", is given by the predicate "x is red" which we apply to them. The first case above is a case of one person having two of these objects, one red, the other square. The second is a case of just one, both red and square. This solution has enormous advantages in explaining various facts about afterimages. It explains why when a person has a red square afterimage, they have

a red afterimage, and, for that matter, why they have a square red afterimage. It explains the various things that are the same and different between three people, one with a red square afterimage, one with a red round afterimage, one with a green square afterimage. It has the advantage of being the beginnings of an account for the impossibility of afterimaging both squarely and roundly when there is only one afterimage: the account is in terms of the impossibility of any thing being both square and round. (Although, on this point, cf Chapter Eleven.) It is amazing, in fact, just how well the picture of real afterimages fits the facts about afterimages (or, to put it in the metalanguage, just how well it fits the entailments between afterimage sentences.)

10. In Conclusion.

The argument I have just presented suffers from the defect that some of its crucial notions need clarifying. For that reason I do not propose to rely on its conclusion in the rest of this thesis. This argument for mental objects (the conclusion was in terms of afterimages, but it is of course intended to have a more general application e.g. to pains) can be regarded as a side-issue in this essay. What I have to say later on will be intended to be neutral (as far as possible, anyway) between adverbial and act-object theories of various mental states. In passing, it should not be thought that the previous argument is intended to work for all mental "entities" for which there exists a substantive. Some mental entities might well be properties, and presumably for them we will not be able to get one or more of its premisses going.

The lesson to be learned from adverbs is clear, I think. It is that linguistic solutions to metaphysical difficulties can be too facile. It has been known at least since "On What There Is" that linguistic items, once their meaning is fixed, carry ontological commitments with their use. It is for this reason that we should be suspicious of locutions introduced "for convenience", in the interests of having some theory-neutral way of describing well-known facts, or for some other metaphysical end - for example "It is as if I see a white horse", "I am illuding a white horse", "It appears to me that there is a white horse", as various ways of talking about hallucinating a white horse. We should immediately want to know what semantic job such locutions are doing (and we should not be fobbed off by pleas that ordinary language is transparently clear). This question of course will plunge us straight into metaphysics. This is not something to be surprised at. Linguistic arguments and metaphysical arguments are not two activities that can be carried on separately, contrary to what seems to have been a widespread belief this century.

So far in this essay, much of the argument has been about linguistic forms, although just insofar as it has been relevant to the mind-body problem, it has had metaphysical implications. In later chapters, specifically in Part Two, the emphasis will shift somewhat, and arguments will take on the appearance of having more to do with extralinguistic fact. Briefly, there are two reasons for this. One is the one we just gave, that linguistic arguments involve us in speculations about the world. The other is something that will emerge from our discussion of Richard Rorty's position. It is, that even if we decide that mental predicates cannot be members

of P, it might be that none are ever true of people. To determine whether they are will involve us in looking at the nature of just what it is about people in virtue of which mental predicates are, or are thought to be, applicable to them. We will now turn to Rorty's arguments.

CHAPTER FIVE, ELIMINATION

1. More General Reduction and More General Elimination.

Our account of the reduction situation so far has been specific to reduction by biconditionals. It might be though that the final theory does not reduce a given theory by means of biconditionals and yet the given theory be true. If the given theory is true, it will have to be dealt with somehow by the final theory (perhaps, for instance, by property identification). This gives a motive for generalising the account of reduction. We introduce a more general reducing formula which expresses the fact that entities of one sort are related in a certain way to entities of another sort. Only the monadic case of this formula is given, for predicates "Mx" in $L(T_2)$, T_2 being the reduced theory, and "Bx" in $L(T_1)$. The general formula for n-adic predicates is more complicated, but it should be clear that it can be exhibited.

The more general formula is "$(x)((Mx \ \& \ \phi_1 x) \rightarrow (Ey)(By \ \& \ \psi_1 y \ \& \ \rho_1 xy)) \ \& \ (x)((Bx \ \& \ \phi_2 x) \rightarrow (Ey)(My \ \& \ \psi_2 y \ \& \ \rho_2 xy))$". "$\phi_1$", "$\psi_1$", "$\rho_1$", "$\phi_2$", "$\psi_2$", "$\rho_2$" are parameters, which in particular cases have predicates for particular properties and relations standing in their places. Suppose that we have a stock of such formulae (we include here the possibility that the more general formulae for n-adic predicates also be in the set) $\{ X_i : i \in I \}$, such that (1) each X_i is true and (2) T_2 is a subset of the closure of $T_1 \cup \{ X_i : i \in I \}$ under deducibility. Then we say that T_2 is reduced to T_1*, and if in addition a predicate $\alpha \in L(T_2)$ is also such

---

* C.f. Nagel 1960 Ch.11.

that $\alpha \notin L(T_1)$, we say that α is eliminated by reduction from T_1.

Some examples will make it clear that this new definition of reduction is a generalisation of reduction by biconditionals.

(1) Taking only the left conjunct, if "x=x" replaces "$\not\phi_1 x$" and "y=y" replaces "$\phi_1 y$", the left conjunct implies "(x)(Mx → (Ey)By)", and we might not wish such a strong relation between Ms and Bs. However, if, for example, "B" replaces "ψ_1" and "F" replaces "ϕ_1", then the left conjunct implies only that if there are Ms which are also Fs, then there are Bs. The function of "ϕ_1" and "ψ_1", that is, is to reduce the scope of the things we are talking about to subsets of { x : Mx } and { x: Bx }. There is also a way of making the formula equivalent to one of its conjuncts. If we let "$\phi_2 x$" be "x≠x", then the right conjunct is valid, and so the whole formula is equivalent to its left conjunct.

(2) An important case is where "ρ_1 xy" is replaced by "x causes y". The possibility that the final theory "reduce" a theory T_2 by giving an account of a causal chain that exists between Ms and Bs, or between two sorts of entities dealt with by T_2 but for whom T_2 does not exhibit the whole of the causal chain connecting them, ought to be allowed for. This last possibility is important since the restriction of reduction to reduction by biconditionals would mean that we could only describe it as reduction if we could make a predicate-by-predicate biconditional reduction of the predicates in T_2, and this seems an unnecessary restriction. This is evidently an argument for describing our generalised conditions as a type of "reduction".

In the rest of the examples, we concern ourselves only with the case where "ϕ_1" and "ϕ_2" are replaced by "x=x", and "ψ_1" and "ψ_2" are replaced by "y=y", i.e. where "ϕ_1", "ψ_1", "ϕ_2", "ψ_2" effectively do not appear.

(3) $\rho_1 = \rho_2 =$ the universal relation (holding between every pair of things). In this case, the formula is equivalent to "(Ex)Mx ≡ (Ex)Bx". Given the previous restriction of concern, this is the weakest relation between Ms and Bs that our formula is capable of expressing, since the left conjunct always implies "(Ex)Mx → (Ex)Bx", and the right vice versa. The same effect is obtained by replacing "ρ_1 xy" by "Mx → By" and "ρ_2 xy" by "Bx → My", which determine stronger relations than the universal relation, but behave equivalently in our formula.

(4) In real life, theories are often tensed, and with the extra machinery of time variables, we might wish to consider relations between things existing or occurring at different times, the causal relation for instance. Our reducing formula allows for this sort of comparison. There are a variety of ways of interpreting quantifiers to allow for time variability. One is to interpret them to mean "there exists at some time" and "for all x at all times". With this interpretation, "(Ex)Mx ≡ (Ex)Bx" means that an M exists somewhen iff a B exists somewhen, which is a very weak relation between Ms and Bs. The only weaker is the null relation, which holds between members of { x : Mx } and { x : Bx } only when at least one of these sets is null.

(5) $\rho_1 = \rho_2$ = identity. In this case, the formula is equivalent to "(x)(Mx ≡ Bx)", our original biconditional. This shows that reduction by means of biconditionals is a

particular case of the more general reduction. Tensing does not of course effect the fact that reduction by means of biconditionals is a particular case. With the above interpretation of quantifiers, "(x)(Mx ≡ Bx)" means that Ms are at all times identical with Bs. If we wanted to make the weaker claim that at the present time Ms are identical with Bs, we could either use quantification over times, machinery for which would exist in the theory, or use a theory without time variables and interpret the existential quantifier to mean the present tense "there exists". There is a useful discussion of these moves in Nicholas Rescher's "Topics in Philosophical Logic".[1]

There is another interesting generalisation that we can make which will lead to a discussion of what has been called "eliminative materialism".

In Chapter Two, we gave a prima facie argument against physicalism. It had two premisses (1) that mental predicates are not members of P, and (2) that mental predicates are sometimes true of some things. Smart and Lewis attempt to deal with this argument by denying the first premise. Some philosophers have seen that another strategy for defending physicalism is to deny (2).[2] If T_2 is the theory of common sense psychology to be dealt with, then we can simply say, with these authors, that T_2, and sentences in it true only when there are Ms (where "Mx" is a suitable mental predicate) are false.

If T_2 is false, we should hope that in the fullness

---

[1] Rescher 1968, Ch. 12.

[2] Feyerabend 1963; Rorty 1965, 1970a; Smart on occasion e.g. Smart 1967. Quine says something quite close to it in Quine 1960. See also Quine 1953.

of time it would be superseded by the final theory. If T_1 is that theory, then T_1 will not of course entail T_2, but the relationships between T_1 and T_2 might be quite interesting. T_1 might entail some subset of T_2, for example, or even none of T_2 at all. It has been argued by Feyerbend that there are many important cases where T_1 supersedes T_2 and there is not even any overlap in the languages of T_1 and T_2. I do not wish to try to decide these issues here, so I will try to make the discussion sufficiently general to allow that any of these approaches may be correct. The main point I want to make is that we can think of T_1 in a very general way as being the theory which is intended to be the right one, and T_2 as having something wrong with it. It might be wrong because it does not cover enough ground, even though it is true as far as it goes, as in normal cases of reduction. Or T_2 might just be false, and superseded somehow by T_1. If any of these things occurs, we will say that T_1 replaces T_2. Replacement is intended to be a vague notion, corresponding to the vagueness in the yet-to-be-precisely characterised idea of supercession. We will further say that, if T_1 replaces T_2, and a predicate $\alpha \in L(T_2)$, and no synonymn of $\alpha \in L(T_1)$, then α is eliminated by T_1. If α is eliminated by the final theory, then we will say that α is eliminable.

A plausible sufficient condition for "Mx" in $L(T_2)$ to be eliminable by the final theory, is if "(Ex)Mx" is false, for, as we pointed out earlier, we think of the final theory as only needing to describe existing things and their properties and relations.[3] The position of the

---

[3]Two comments are needed. The first is that

above philosophers who deny that troublesome mental predicates are ever true of anything, can be (partly) characterised by saying that they hold that mental predicates are eliminable.[4] We will call this latter view "eliminative materialism".

The version of eliminative materialism we will consider in this chapter, is eliminative materialism coupled with the thesis that sentences of the form "(Ex)Mx" where "Mx" is a (troublesome) mental predicate, are never

---

this is not a necessary condition for, as we shall see, contingent property identification gives a way of saying both that "(Ex)Mx" is true, and also that "Mx" is eliminable. The second comment is that the above is not quite a sufficient condition. Laws in the final theory might contain machinery for dealing with variables certain values of which are never realised e.g. very high temperatures. It seems to me that the case in hand, the eliminability or otherwise of mental predicates, is not like this, and we will proceed to ignore this complication. Certainly it would be unreasonable to expect the final theory to contain machinery for denying the existence of whatever, no matter how fantastic, does not exist. Notice, by the way, that eliminability does not imply meaninglessness. We can quite reasonably use an eliminated predicate to deny existence.

[4]It is characteristic of the philosophers mentioned that they are inclined to deny the analytic-synthetic distinction and with it synonymy. Our definition of elimination ought to be acceptable to them, however, because from the assumption that a predicate has no synonyms, it follows that it has no synonyms in $L(T_1)$. It is difficult to see, moreover, what force can be given to Quine's recommendation that mental terms be "dropped" without some such notion. Quine can hardly be too concerned to drop mental predicates considered as syntactic items, or even as items with a certain extension. In any case, this is a problem about whether the views of certain philosophers can be made to fit our definition, not a problem about the definition itself, though it might have its problems.

true. This theory evidently has considerable advantages, in particular it avoids the necessity for showing that mental predicates or theories are topic neutral. Rorty, and on occasion Smart, have expressed the opinion that mental language is most likely loaded in favour of dualism.[5] Also, it can avoid the knotty problems involved in the questions of property identification, and whether biconditional reduction can ever be sufficient for elimination.

The main drawback of this version of eliminative materialism is without doubt that it has to deny that predicates like "x has a yellow afterimage" and "x has a stabbing pain" are ever true of anyone. That is a hard thing to accept. As Smart and Place both pointed out quite early on in the debate as it has been conducted recently, "I have a yellow afterimage" does seem to function as a genuine report about some state of people, a state which people sometimes do have. Richard Rorty[6] has noticed that this sort of elimination, denial that any troublesome mental predicate is ever true of a person, must at least be accompanied by an explanation of how the (allegedly) false sentence "I have a yellow afterimage" can apparently be used to report something about the speaker. If "I have a yellow afterimage" is always false, it is nevertheless sometimes appropriately used. There is something right about its use.

2. Rorty On Elimination.

In what follows, we will be concentrating mainly on Rorty's version of eliminative materialism. Rorty

---

[5] Rorty 1965. Smart 1967.

[6] Rorty 1965.

points out that there are perfectly respectable senses of the verbs "to call", "to report", "to describe", in which we can say e.g. that what used to be called "a certain quantity of caloric fluid" is now called "a certain mean kinetic energy of a certain group of molecules", or that what used to be described as "Zeus' thunderbolts" are discharges of static electricity, or that when people used to report possession by demons, they were really reporting their hallucinations.

Rorty describes his theory as a variant of the Identity Theory, but he makes it clear that the sense in which mental states are identical with brain processes is not strict identity, but rather "roughly the sort of relation that holds between existent entities and non-existent entities when reference to the latter once served (some of) the purposes presently served by reference to the former."[7] Rorty's "mistake" here is easily explicable by the fact that innovators are often not aware of the full significance of their discoveries. Rorty's basic formula for the relation does not use the verb "to refer", but the verb "to call". (He does use these other verbs sometimes, as variants.) The sense in which X is identical to Y is that expressed by the paraphrase "what people used to call (or in some cases, now call) 'X', is Y". Another example is the sometime use of the predicate "x is a unicorn horn". In the early days of whaling, numbers of narwhal horns used to be found, and it was thought that they were unicorn horns. There are no unicorns, so any sentence which entails that there are unicorn horns is false. However, someone could have said "I was present at the discovery of the biggest unicorn horn ever found",

---

[7] Rorty 1965 p.175.

or "Have you seen Captain Ahab's new ornament? He's painted that unicorn horn he found tangerine", and there have been a real object which was being (mistakenly) called a unicorn horn. And so it is with mental predicates, according to Rorty. Mental language contains a false theory, but we can and sometimes do, use mental language to talk about genuine goings-on, neurological ones.

Rorty's examples seem pretty clearly correct. Donellan has recently drawn attention to the fact that successful reference can be achieved even though the description intended to identify the thing in question is not true of it.[8] Furthermore, I do not think that we need to restrict ourselves to the "call" formula. In one central sense of "report", one reports events by asserting their occurrence. But with an ontology of events and machinery for reference to them, one can just as much call a particular event the wrong thing as one can an object: Captain Ahab's painting of his unicorn horn occurred at t. Now such sentences have an equivalent form which does not contain a singular term for an event: "Captain Ahab painted his unicorn horn at t". So if the first reports an event, it is easy enough to say that the second does too, even though there is no explicit reference

---

[8]Donellan 1972. One of Donellan's examples is of meeting at a party a man whom I mistakenly take to be J.L. Aston-Martin, the well-known author of "Other Bodies". I might falsely report to my friends later that I met Aston-Martin at the party, but later on in the conversation successfully refer to the person I did meet: "And then Aston-Martin punched Robinson ...". This is an example of reference using a proper name, but I do not think much hangs on the difference between proper names and descriptions here.

to events in it. Thus we can say that what the second reports or what a person reports by using it is identical with a certain event (the painting of a unicorn horn) even though the sentence is strictly false.

What we can conclude from these points, I think, is that the flat denial that e.g. anyone ever has afterimages has some of the sting taken out of it. If "I have a yellow afterimage" is sometimes correctly used to report some state of myself (which is in fact a neurological state), then while I might have been mistaken as to its nature, Rorty's theory is not so counterintuitive as to imply that I was wrong in thinking that I was in <u>some</u> special state, on those occasions when I uttered the sentence.

One point I would like to stress is that we have already introduced some re-interpretation of Rorty's position here. Rorty denies that there are sensations, and this is not quite the same as denying that anybody ever has sensations. An adverbialist, for example, might be happy to deny the existence of sensations, afterimages, pains, but not want to deny that anybody ever has them, because he or she thinks that the semantics of "x has a sensation" does not include quantification over mental objects. I do not know whether this distinction ever occurred to Rorty; there is no evidence that I can find in his writings that it did. However, what Rorty says applies at least as well to the denial of the having of sensations as it does to the denial of sensations, and furthermore interpreting Rorty in the former way makes his position an interesting one. So I will proceed to interpret Rorty as denying that anyone ever has sensations, and, more generally, as denying that troublesome mental predicates are ever true of anything.

Rorty's position has received considerable criticism. I will argue that with certain further plausible re-interpretations it can resist those criticisms.

3. In Defence of Rorty.

In addition to holding that mental predicates are not true of anything, and that they are used to report events which are neurological events, Rorty also seems to want to hold that human beings do not hold any false beliefs about their mental states. In one place (p.182), he says that he "does not wish to say that people who have reported sensations in the past have (necessarily) any empirically disconfirmed beliefs". Not much hangs on the "empirically disconfirmed", I believe: on p.181-2, for instance, he suggests that his view is similar in this respect to a view about the relation between "x is a table" and "x is a cloud of molecules", the use of the former of which "does not suggest or require as a ground that people who say 'This is a table' hold false beliefs".[9]

Now this is inconsistent with the view that we have been attributing to Rorty. There are two ways of resolving this inconsistency. The first way is to interpret Rorty as saying that the properties expressed by the predicates "x has a sensation", "x has a yellow afterimage" and "x is a table" are (contingently) identical with certain physical properties. If we hold this, then we can, plausibly, hold two further propositions: (1) that "x has a yellow afterimage" is eliminable, since in the final theory the same facts about what exists and what

---

[9] Rorty 1965 p.182.

its properties are can be expressed by sentences of the form "a has F" where "Fx" is a predicate which expresses the same property as "x has a yellow afterimage", (2) "a has a yellow afterimage" is sometimes true (not false), since the property which the predicate expresses is sometimes possessed by some things. This sort of interpretation certainly fits with Rorty's above statements about false beliefs, but it does not square with the denial that mental predicates are true of anything. We will be discussing this version of eliminative materialism in the next chapter, and we will not pursue it here.

A way of resolving the inconsistency which is closer to the spirit of Rorty, I think, is to say that he is just wrong in saying that humans do not have false beliefs when they say that they have afterimages. If "(Ex)(x has a yellow afterimage)" is always false, and if humans do sometimes believe that they and others sometimes are having yellow afterimages, then those beliefs are false, and moreover the predicate in question is eliminable. In what follows, we re-interpret Rorty (for a second time) so as to make his position consistent in this way.

A word about beliefs is necessary. In admitting beliefs, Rorty might seem to be contradicting eliminative materialism, for surely "x believes that x has a yellow afterimage" is a troublesome mental predicate. But this remains to be seen. Beliefs might look prima facie to be a trouble for physicalists, but some physicalists, e.g. Place [10], have thought that they are no problem in that

---

[10] Place 1956.

Sensations are Brain Processes	'Sensation' denotes	There are sensations	We have certain false beliefs	CR	OR	
Yes	Yes	Yes	No	Yes	No	RM.
No	Yes	No	No	Yes	No	WEM.
No	No	No	Yes	Yes	Yes	SEM

Fig. 1. Lycan and Pappas' classification.

they have a straightforward dispositional analysis. In - sofar as our principal concern in this book is with predicates like "x has a yellow afterimage", rather than belief predicates, I see no reason not to tolerate belief predicates. There is certainly no inconsistency in being eliminationist about afterimage predicates and dealing with belief predicates in a different way. If someone even denies that we have any beliefs about our mental states, indeed that we have any beliefs at all, then I do not know how to deal with them, for they seem to be denying something that is obviously true.

4. Lycan and Pappas' Criticism.

With the above modification of Rorty's view, the principal criticism that Lycan and Pappas [11] make of it becomes irrelevant. In order to understand their criticism, it is necessary to reproduce their schema for classification of various materialisms. (See Fig.1.)

RM is Reductive Materialism, Lycan and Pappas' version of the Identity Theory. WEM, Weak Eliminative Materialism, according to them, is Rorty's position. SEM, Strong Eliminative Materialism, a stronger eliminative position than Rorty's. CR is the thesis subscribed to by Rorty in Rorty 1972, that "at no greater cost than an inconvenient linguistic reform we could drop (mentalistic) terms". OR is the stronger thesis that we "ought to drop mentalistic terms". In addition, according to Lycan and Pappas, all materialisms subscribe to a thesis M: "if x is a sensation, x is a brain process".

---

[11] Lycan and Pappas 1972.

Lycan and Pappas present a dilemma designed to show that WEM collapses either into RM or into SEM. Either "x is a sensation" entails "x is not a brain process", or it does not. Suppose it does. Then, since WEM is a materialism, from M we can conclude that there are no sensations. It follows from this that we have certain false beliefs, which contradicts WEM, and which is the "earmark of SEM" (p.155).

This argument points to the difficulty about false beliefs that we have just raised. But it is doubtful whether having no false beliefs about certain of our mental states is the "earmark" of Rorty's position. Rorty defines his position with respect to the claim that mental terms do have a referring/denoting function. The earmark is very much a fleabite - in fact I will later argue that Rorty elsewhere commits himself to denying it - and one which we have urged should be dispensed with. The dilemma fails, then, because the first horn fails.

It is worth noting that Lycan and Pappas attribute to the Identity Theory the thesis that mental terms could be dispensed with (i.e. CR above). But, as we have seen, Identity theorists who employ the device of topic neutral analysis are committed to saying that mental predicates are members of P, and, far from being eliminable, are a necessary part of our causal description of the world. Indeed, it is precisely the thesis of eliminability that distinguishes Rorty's position from Smart's version of IT.

5. Cornman's and Bernstein's Objection.

There is a more interesting objection to Rorty,

due to Cornman[12] and Bernstein[13], which will cause us to modify what Rorty says even further.

Cornman and Bernstein both make a distinction between observational terms and theoretical terms.[14] Of theoretical terms, they both say that they can always be discovered to be dispensable because we will never be in a position to know that a given theory is the correct one. We can never eliminate observational terms, on the other hand, for they stand for what we are aware of, or experience. An attempt to eliminate them by replacing them (in the case of mental talk) with neurological predicates would only have the effect of the neurological predicates changing their meaning so that they entailed what they do not entail now: that there are sensations. Cornman argues that mental predicates are not synonymous with any neurophysiological predicates,[15] and that therefore they perform a descriptive function over and above that of neurophysiological predicates: there are things that they describe - better, there are properties that they express - that are not expressed by any neurophysiological predicates. So we cannot say that we can dispense with mental predicates without replacing them with something which does entail that they are satisfied. As Rorty sums it up, "What we are aware of is not postulated, and only the postulated is eliminable."[16]

---

[12] Cornman 1968a, 1968, 1971.

[13] Bernstein 1968.

[14] This needs to be slightly qualified: Cornman's distinction is more complex than this, but the complexity does not matter for this argument.

[15] Cornman 1968-9.

[16] Rorty 1970a p.227.

Rorty chooses to defend his position by attacking the distinction between theoretical and observational terms, and with it the idea that there is a class of "observational" predicates which are ineliminable.[17] His reason for thinking that the distinction is illegitimate, is that he holds that there is no such thing as something that "appears to us, or what we experience, or what we are aware of"[18] independent of the language we use. If we spoke a different language, then our experiences would be different. "If we got in the habit of using neurological terms in place of 'intense', 'sharp', and 'throbbing', then our experience would be of things having those neurological properties and not of anything (eg) intense." (p.228) "If it were the case that we experienced the same thing when we used the new vocabulary as when we used the old, then their point would be sound. But there is nothing to be this same thing." (p.228) "... there is nothing in common between the two experiences save that they are had under the same conditions." (p.228)

What can Rorty mean by this? He seems to be saying quite literally that if we talked brain process talk, we would be aware, in introspection, of brain processes, and if we talked sensation talk (as now), then we would be aware of sensations (which ex hypothesi are not identical with brain processes). This is a very strange suggestion. If anything, it would have as a consequence that we ought not to eliminate our sensation vocabulary, lest we lose our minds.

Something else he might mean is that in intro-

---

[17] Rorty 1970a.

[18] Rorty 1970a pp.227-8.

spection we are aware of brain processes ~~all the time~~; there is nothing dualist about the things we are introspectively aware of. But when we come to talk brain process talk, we come to be conscious that our inner states are physical states.

Rorty does not give a reason for thinking that this is true. In another place[19] he endorses the positions (attributed to Feyerabend, Sellars and Kuhn) that one cannot separate any theoretical component of sentence verified more-or-less directly by observation from an "experiential" component. He also says that the reason why this is so is that one cannot separate any conceptual component of the mental state we are in from a "pure sensory" component, and that furthermore, one cannot separate any observational component from a component which is a consequence of our background theory. Perhaps this is a reason for thinking that our experiences would be different if we talked brain process language. It is not a reason for thinking that as things stand at present, we are aware of or experience only neural items. This is a crucial point about Rorty's view. Another way of puttingit is to say that Rorty might be right in thinking that using the brain process vocabulary would change our inner life, but it will not follow from this that our new set of beliefs about our inner states were not as wrong as our present ones must be if Rorty is to be correct. The new vocabulary might be impoverished and we not know it. Unless Rorty gives an independent reason for supposing that we are at present only aware of physical items when we introspect, there is no reason to believe

---

[19] Rorty 1972.

what he says.

We are a little ahead of ourselves here, and we will return to this point later. What I want to make clear is how Rorty can answer at least part of Cornman's objection. He tried to answer it by attacking the theoretical-observational distinction. An easier way, and one to which Rorty elsewhere commits himself, is to deny that mental states are incorrigible. I say "easier", and readers might find it much harder. For that reason, I will presently digress to argue for the corrigibility of mental states. Let us note, though, that Cornman's point depends on the claim that we are aware of having mental items, say sensations. Now certainly Rorty denies that we have sensations. It is also undeniably true that most of us believe that we have them. Let us then just say that those beliefs are false. It is a conclusion which we arrived at previously by another route anyway.

Let us digress and look at incorrigibility. An argument for the corrigibility of our beliefs about our minds will evidently give backing to Rorty's position.

6. Incorrigibility.

There are three separate "incorrigibility" theses which I believe to be false.

(1) Necessarily, if x believes that he or she is M, then x is M.

(2) Necessarily, if x is M, then x believes that he or she is M.

(3) Necessarily, if x is M, then it is not the case that x believes that he or she is not M.*

(3) is somewhat harder to argue against than (1) or (2), though I believe it to be false. I will concentrate on (1)

* where 'M' is a mental predicate.

and (2).

Now there are some cases of "M" where few people would dispute a corrigibility thesis, for example, emotional states such as jealousy. The hard cases for a corrigibility theorist are cases like pains, afterimages, hallucinations, seeing red and so on. I will concentrate on the hard cases.

Why should we believe (1) true? A common answer, and one which Frank Jackson gives in a recent article, is that (1) is analytic.[20] Immediately we have an impasse, for it is notoriously difficult to resolve disputes over the analyticity of a claim. Counterexamples to an analyticity claim are typically met by redescribing the example so that it fits the analyticity claim. I wish, however, to try to give some sorts of reasons for thinking that (1) is false. I will say straight away that I do not think that I can <u>prove</u> (1) (or (2) or (3)) false against a determined defender of it, for just the above reason. Sometimes a dispute about analyticity comes down to one side whiteanting the resolve of the other to use words in just his or her way.

Let us ask: who would the onus be on to prove their case, the affirmer or (1) or the denier of (1)? I must say that I think that the onus is on the affirmer. I do not know how to prove this, however, so I will not rely on it. A position which does seem to me more clearly defensible, is that the onus is on both the affirmers and deniers of (1) to prove their respective cases, in the absence of which proofs we should be agnostic about (1). This follows from the general principle that we should

---

[20] Jackson 1973.

suspend judgement in the absence of reason to believe. Sometimes the onus in a situation can be moved by some such argument as that one of the alternatives is simpler (as in arguments about the rationality of belief in God). It is my suspicion that positions with less analyticities in them are generally conceptually more economical. That is why I believe that the onus is on the defender of (1) and not on the denier of (1). I do not know how to show that this suspicion is true, so my belief about where the onus lies is a weakly held one. Anyone who feels that the suspicion is justified will have an additional reason (over and above the ones I will try to give here) for accepting my conclusions.

Let us start, then, from the agnostic's position. Jackson thinks that the failure of arguments for denying (1) is enough to swing the onus away from the affirmer of (1) and onto the denier of (1). If he is right, then our starting point is an incorrect one. But his being right depends on there being no good arguments for denying (1), and that is what is disputed here.

The first argument I will give is a version of the well known "distinct existences" argument.[21] Let us suppose that the mental state we are dealing with is pain. The argument is intended to go through for a variety of states e.g. afterimaging. The first step in the argument is to establish that in any case where a person both has a pain and believes that they have a pain, the belief and the pain are distinct.

---

[21] Cf. Smart 1963a, Armstrong 1968. Another interesting argument for corrigibility which we will not discuss here is given by John Chandler 1970, and F. Verges 1974.

By "distinct" I mean "numerically distinct". I claim that the pain I have at t and the belief at t that I am currently having a pain, are not the same thing.

The belief that I have a pain is distinct from my pain, because the former does not exhaust the latter. There is more to having a pain than believing that you have it. Pains or their havings have qualities over and above the beliefs about them, we might say. If this were not so, then there would be no way of distinguishing two different sorts of pains, a burn and a stab say, from one another. All the differences between the two would also be there in the difference between believing that you have a burning pain and believing that you have a stabbing pain. But if we cannot distinguish these by distinguishing between the (believed) natures of what is had, we cannot distinguish them at all. Thus to distinguish between the natures of the pains is to distinguish between qualities beyond those of the beliefs.

The next step in the argument as Armstrong and Smart give it is to show that if pain $\neq$ belief about pain, then one can exist without the other. They invoke Hume's dictum that if x is distinct from y, we can always conceive of the one existing in the absence of the other. It is easy to show that this is false, as Jackson points out. Husbands are distinct from wives, but "I am a husband" entails "I have a wife". A husband cannot exist in the absence of a wife (i.e. when his wife does not exist), for then he would not be a husband. It is similar with pains and beliefs. It might be the case that they are distinct but nevertheless, like husbands and wives, one cannot exist without the other.

Here, however, we can find something strong

enough for my purposes. If a wife ceases to exist, so does a husband. But not, of course, in the sense that the husband dies too. If John's wife dies, John does not necessarily go out of existence. Barring polygamy, all that necessarily happens is that John ceases to be a husband. Husbands and wives are logically linked, but only by virtue of bearing those descriptions. Nothing logically prevents something like John from existing, exactly like him in every respect save that he is not a husband. Similarly, even if it were the case that as a matter of logical necessity pains and beliefs about pains went together, this could not prevent something existing exactly like a pain save that it was not accompanied by a belief, and something existing exactly like a belief about a pain save that it was not accompanied by a pain.

This is a standard enough point to make in disputes about whether something is analytic. Someone denying e.g. the analyticity of "All swans are white" can sometimes get their opponent to admit the possibility that something exists exactly like a swan in all essential respects save that it is black.

There is one twist in this instance. It might be replied that as far as the matter of pains and pain-beliefs is concerned, the above formal point does not show that if you took the belief away there would be anything left to be "exactly like a pain save that it was not accompanied by a belief".

That is why I have prepared the way by arguing that there is something to pains over and above the belief. For if that is true, then if the belief did not exist, then there could still be something existing exactly like the pain except that it was not accompanied by a belief.

Similarly, because there is something to beliefs about pains other than the pains themselves, we can say that something exactly like a belief that we have a pain can exist even when the pain does not, exactly like the belief except that it is not accompanied by the pain.

Having established this much, I want to say that I apparently use "pain" differently from Jackson. I use "pain" to denote those features of the pain/belief situation that are present when the belief is present or absent. And I use "belief about pain" to denote those features common to the situation in question when the pain is present, and when it is absent. Do not confuse this with saying that I use "belief that I have a pain" to denote irrelevancies like my heartbeat, which are present when I have a pain and when I have not. I am not offering a definition of "pain" and "belief" here. It is presupposed in this discussion that we can tell our pains and beliefs from our heartbeats.

I cannot think of a good reason for Jackson's usage being the correct one. (Remember that our starting point was the agnostic's position.) I can think of a reason why Jackson might use words the way he does. He might be unable to imagine what it would be like for he himself to believe that he had a pain and not have one, and vice versa. Evidently this would not be a good reason. We might be unable to imagine what it would be like for ourselves to be five-dimensional creatures in a five-dimensional space, but this does not mean that space could not be five dimensional. Nevertheless, such an inability could lead to an unwillingness to accept the conclusions being urged. A case would clearly be more satisfactory in this respect. The trouble

with giving cases, however, is that, as we have noted before, cases can always be redescribed so as to preserve analyticities. We might add that this procedure is not always illegitimate.

Cases can, however, sometimes serve to incline without necessitating. So the second argument against incorrigibility consists in giving cases. I think the best cases of having a sensation and failing to believe that we have are those surrounding the phenomenon of attention. Making automatic avoidance behaviour of obstacles on the road without being aware of doing it or of the obstacles and later without being able to remember anything at all about the situation. Exhibiting pain behaviour (wincing, shifting the weight, etc.) without being conscious of a pain because the attention is on something else (or been because we are asleep). Being affected by subliminal advertising. There is no doubt that in such cases some unconscious state of ours occurs which plays the same sort of causal role in our behaviour as do pains and visual experiences in more normal situations.[22] Why should we call those unconscious states pains and visual experiences? The similarity in their causal roles is one reason. Another strongly suggestive reason is that we can come to learn to be aware of sensations in such subliminal cases; and that what we come to be aware of is what it is that plays the causal role that these unconscious states hitherto played. Redescribing this state of affairs so as to preserve the analyticity would seem to require us to say that while learning the sensation gradually takes over the causal role of an entirely different unconscious state - different in that it is not a sensation - instead of saying

---

[22]It needs to be established that there are such things as visual experiences in normal cases of perception. This is argued for in Ch. Nine, Ten.

that it is the same sort of thing that we gradually acquire the ability to become aware of, and that we are sometimes conscious of, and at other times have no beliefs about.

In this connection, it is appropriate to mention Jackson's treatment of the speckled hen example. We will treat the speckled hen example in greater detail later[23], for it is particularly important in connection with the philosophy of perception.

> If I look at a speckled hen, no doubt it will appear to have more than ten speckles; but there will be a number, depending of course on the particular hen, which will be such that I hesitate, indeed am unable, to say whether the hen appears to have more or less than this number of speckles. The obvious explanation of this is not that the hen looks to have a definite number of speckles which I am unable to specify, but that the hen does not look to have a definite number of speckles at all. [24]

Jackson might just be making a point about the logic of the word "looks". If so, he has sold the example very short, as we will see. Imagine not looking at a speckled hen, but hallucinating one, and not knowing the number of speckles. We make the following assumptions, all of which will be argued for when we treat the example more fully: hallucinating a speckled hen is having a complex of visual sensations, which are not just beliefs or suppressed inclinations to believe (as for Armstrong and Pitcher), and its speckles are features of those sensations. The example is not exactly the same as Jackson's, but close enough to make it reasonable to think that his reply to it will be the same. His reply, then, will be that the (hallucinated) hen does not have a

---

[23] Together with a more full discussion of attention.

[24] Jackson 1973 p.60.

definite number of speckles. That, I claim, is implausible. It is implausible because we can imagine counting the speckles on the hallucinated hen and getting a definite answer, say fifty. But it is unreasonable to think that the hen necessarily must change during the counting process, from having an indeterminate number of speckles, to having fifty. Surely we can "keep an eye on" the whole collection of speckles while counting, to ensure that they do not change: there was the same number before the counting as after, and that was fifty. There was not an indeterminate number before counting, but fifty, and we had no beliefs about that fact. Thus, our sensations can have some features which we do not believe them to have.

Reverse cases are harder to find i.e. cases of beliefs that we have sensations in the absence of the sensations. A quick case is where we make a mistake in counting the number of speckles on the hallucinatory hen, and so come to believe falsely that the hen has fifty-one speckles. A more detailed case is as follows.

We go to the dentist with a sore tooth. He anaesthetises the tooth and it stops hurting. He tests the gum to see that the anaesthetic has taken effect, and it has. Our attention is somehow distracted and we do not see him test the gum; we are very nervous and if we had seen that we would have been very frightened. He brings the drill down to within 1mm. of the tooth and suddenly we see it. It does not reach the tooth but our nervous anticipation is such that we suddenly believe that it has reached the tooth and that it hurts very much. We leap up from the chair and sprint across the room to the door, with the dentist in hot pursuit, the dedication of a true healer in his eyes.

At this point the case splits into two subcases. In one, he fails to catch us. In the other, he catches us. In the first we escape down the street, convinced that he is a butcher, and proceed to tell our friends so. In the second, he leads us back to the chair. After some remonstration, we become calm enough to remember back that it did not really hurt, but that we only thought it did.

Jackson's reply to a similar case given by Don Locke[25] is to point out that the description of the case does not entail that we are mistaken in believing that we have a pain. What happens might be that we really do have a pain, caused by the sight of the drill in our mouths. The description that Jackson favours is that we believe propositions (e.g. that we are touched by a drill, etc.) which entail that we have a pain, but do not believe that we have a pain.

At the risk of boring the reader, we repeat that it is a matter of refuting an alleged analyticity and that is always difficult. What will we say of the case? There is no doubt that immediately afterwards we believe that we had a pain. Perhaps this belief starts instantaneously the drill comes out of the mouth, and until then we believe only that we have a sensation of shock. It seems strange to fix a time to the beginning of the pain belief like this. I cannot think of any reason for describing it Jackson's way save adherence to the analyticity. It certainly seems reasonable to say that we did not have a pain if we can in calmer moments remember back that this was so.

---

[25] Locke 1967.

What would it feel like to be in this state? Well, in your agitation and confusion you have momentarily the feeling of certainty that it is hurting, but at the same time all there is is the feeling of shock, which causes you to have this belief. You confuse your shock for pain. I am convinced that I can imagine this, and that it is a genuine case of mistaking shock for pain.

If belief can occur without pain, why should it be so hard to imagine? Why should not cases occur more often? I suspect that the difficulty is only a philosopher's difficulty. Once we begin to think that there are cases of belief without pain, it becomes easier to believe that much more common cases are also cases of belief without pain; for example the hypochondriac who characteristically mistakes a feeling of pressure in the bowels for pain.

I think that the two arguments offered here for corrigibility lend support to one another. Accepting the Distinct Existences argument should incline us to be more favourably disposed to the cases, and the cases make it more reasonable to believe the premisses of the Distinct Existences argument that belief does not exhaust pain, and that words can properly be used differently from the way Jackson seems to use them. I conclude that the arguments given make it reasonable to believe that both the incorrigibility theses (1) and (2) above, are false.

7. Rorty on Incorrigibility.

This completes our digression into the question of the incorrigibility of our mental beliefs. The conclusion backs up an eliminationist who wants to deny that mental predicates are ever true of anyone. He or she can say that our beliefs that we do have mental states are false.

It backs up Rorty against Cornman's objection. Rorty can say that nothing is in principle ineliminable, and he can certainly say that the grounds that Cornman gives for the ineliminability of sensation predicates - that we know we have sensations - are simply false.

This defence to Cornman is one Rorty does not offer. Perhaps Rorty did not see it because in another place he defends what he describes as an incorrigibility thesis.[26] However, the incorrigibility thesis he wished to defend was the thesis that mental states are incorrigible only in the sense that there are no accepted criteria for overruling first person reports. Now "accepted" is a time relative word. Something can become accepted at a time later than when it was not accepted. We could, consistent with this claim, come to discover ways of determining that our introspective beliefs were false. As Rorty himself notes, that mental states are incorrigible in this sense is consistent with the idea that mental states are corrigible in another sense, namely the sense that one's beliefs about one's mental states can just be wrong.

Furthermore, Rorty spends section 5 of Rorty 1970b arguing just this latter claim. It is difficult to see, then, why he did not use it as a defence against Cornman. In passing, it seems odd that Rorty should want to defend incorrigibility in even the weak sense that he does. For if there are no accepted criteria for overruling introspective reports, and some of them are reports of having sensations, which not-accepted criteria does Rorty employ in deciding that no-one has sensations?

Denying incorrigibility is an answer to this

---

[26] Rorty 1970b.

objection of Cornman's then. It has as a consequence that human beings have been systematically making a mistake in thinking they have sensations. A new problem then arises: what explains the systematicness of the mistake?

Notice that it is not such a good answer to talk about the language we speak determining the experiences we have. That, as we said before, is consistent with our having experiences not of brain events but of dualist events. It must be supplemented with the claim that we have always been wrong in believing that we have sensations, and then we will want to know why always?

Asking "why always?" is not a knockdown refutation of Rorty's position: If a mistake about our mental states is ever possible, I suppose it is possible always.[27] Asking "why always?" does however point up one danger in the Rorty position: we should be extremely wary of a prodigal use of the corrigibility thesis.

After all, if dualism were true, and we did experience items as having properties which were not the properties of anything physical, a cavalier use of the corrigibility thesis would prevent us from ever advancing this fact as an argument against materialism. That is to say, a cavalier use of the corrigibility thesis could be used against any introspective evidence against materialism. One could just not countenance any refutations. That is surely not respectable. We would wonder, if this move were allowable, what all the fuss

---

[27] This must not be confused with saying that all our beliefs about everything might be wrong.

has been about. Surely one of the principal problems with materialism has been to account for introspection. If we are just unmoved by it, materialism is a very easy thesis indeed.

8. Conclusion.

Rorty's position has been considerably modified in this discussion. We have modified it in three ways. We have modified it so as to make it a more rigorous eliminationist position, rather than one which merely denies sensations. We have modified it to cater for the problem of false beliefs, by taking the position to be that human beliefs about having mental states are false. We have modified it by denying the incorrigibility of mental states to back up the second modification. So modified, I believe that it is a strong position.

In Chapter Two, it was remarked that this sort of position might be one way of interpreting Smart's words in Philosophy and Scientific Realism : that the topic neutral analysis, while not strictly an analysis, purports to give in a general way what sensation reports are about. Smart might have said that while there is no analysis, and so mental predicates, literally anyway, are eliminable,(because being loaded with dualism they are not true of anyone) still mental terms have a reporting function. All that goes on that we are introspectively aware of, he might have said, are the occurrences of similarities between mental states, causal relations between states, and between states and the world. Mental terms, while not strictly true of the items of which we are introspectively aware, can be used to report, denote, refer to them.

So interpreted, in fact, Smart's position has an advantage which Rorty's lacks. For Rorty does not give any more of an account of what we are introspectively aware, than to say that we believe that we have sensations and that those are false beliefs. He does not, for example, show why it is that mental terms do denote brain processes rather than nothing at all. This connects with the question: why have human beings made such a systematic mistake about their minds if Rorty is right? We might expect that human beings can discover something about their states on certain occasions, e.g. when they look at a coloured square in a psychology text and then look at a blank page. And surely it is reasonable to think that those features that they do discern are what mental terms denote. A dualist, for example, will characteristically say that human beings discern non-physical features on those occasions. If humans do not discern those features, surely we are owed an account of what they do discern, if only for the reason that it might turn out that the dualist is right. After all, it is unlikely (to say the least !) that humans are aware of their pains as brain states.

In fact, this is where all the trouble comes from about the physicalist status of mental predicates. Humans are aware of something on those occasions, and they use mental predicates to report what they are aware of. If we just say that humans are aware of nothing, the systematicness of the mistake becomes overwhelming. We should want to know why we feel inclined to use those mental predicates at all, and why we should want to use them to report anything at all.

If Rorty's view is deficient in this respect, Smart's reinterpreted is not. At least it has something to say about what we are aware: we know about the similarities and causal relations of our states, and we use mental predicates to report these features. It even contains the beginnings of an explanation of the systematicness of the alleged mistake. Introspection makes us aware only of very general and topic neutral features of our states, and these features could be the features of physical things or nonphysical things. Introspection does not reveal our mind to be physical.

I do not want to pursue the possibility of making this explanation satisfactory. The discussion does, however, point up one important thing. Smart might be right in claiming that this is all we know about our mental states, but that remains to be seen. It might be possible to show that what we know about our minds are facts which are inconsistent with physicalism. In fact, we will be arguing for just this conclusion in Part Two. Indeed, surely just this sort of investigation is forced on us by the consideration that the reason that mental predicates are troublesome has something to do with the phenomena revealed by introspection. The introspection situation is precisely from where many mental predicates derive one of their principal uses. We ought, therefore, to look at what we know and can come to know about our mental states. If Smart and Rorty are right, then we should not be able to find out that what we know are propositions inconsistent with physicalism. Contrapositively, if we <u>can</u> find out such facts, then Smart's account of what we introspectively know is an incorrect one, and furthermore, Rorty cannot be right in claiming that mental terms denote only brain

processes.

Discussion of these points will be deferred for some time, however, as we have not yet explored all the avenues of defence open to the physicalist. In the next chapter, we will look at a different attempt to show that mental predicates are eliminable. We will now turn to this.

CHAPTER SIX. ELIMINATION WITHOUT IMPOVERISHMENT: CONTINGENT PROPERTY IDENTIFICATION

1. Another Kind of Eliminative Materialism.

A predicate is eliminable when it does not appear in the final theory. In the previous chapter, we discussed a type of eliminative materialism which depended on accepting a certain sufficient condition for the eliminability of various mental predicates (say "Mx"). The condition was that "(Ex)Mx" be false, for in such a case "Mx" would not be needed to describe which things exist and which properties and relations they have.

In this chapter, we will discuss another version of eliminative materialism. According to this version, the sentence "(Ex)Mx" might be true in the sense that the property expressed by "Mx" is possessed by something, and yet "Mx" still be eliminable. Now from the definition of eliminability, this could occur only if no synonym of "Mx" occurs in the final theory. But if "(Ex)Mx" is true, we should need some way of saying the same thing that "(Ex)Mx" says, if the final theory is going to omit nothing. We should need another way of expressing the fact that the property which "Mx" expresses is instantiated, and we should need to be able to do it without using a predicate synonymous with "Mx". For this to be possible, we will need it to be possible that nonsynonymous predicates express the same property. Conversely, if "Mx" and "Bx" express the same property, then it seems clear that one of them will be unnecessary in the final theory, and so can be dropped (providing that it does not have a synonym remaining in the theory, of course).

So this version of eliminative materialism claims that mental properties are in fact identical with physical properties, even though terms that express, or name, those mental properties are not in any sense analysable into physicalist predicates or property names. Because mental predicates express properties which can be perfectly well dealt with by using physicalist terms, they can be dropped. However, unlike with Rorty's theory as we interpreted it in the last chapter, the theory does not have to deny that there is at least a sense in which it is true that e.g. people have afterimages. That is an obvious advantage over Rorty's view.

It must be said that this sort of theory might have been what Rorty had in mind when he denied that people's beliefs about their mental states were false. If we construe Rorty this way, then it certainly makes sense of that denial. On the other hand, it would be difficult to reconcile this interpretation with Rorty's denial of sensations (or, as we have modified it, with Rorty's denial of the instantiation of mental predicates).

It might be wondered why one should bother eliminating the predicate "Mx" if it expresses the same property as a predicate "Bx" in the final theory. Surely to retain it could do no harm to physicalism. Recall that our definition of "eliminability" was that the predicate in question does not occur in the language of the final theory. Now why, goes the question, hold that "Mx" is eliminable at all in the circumstance where it expresses a property expressed by a predicate in the final theory?

The answer* lies in the fact that the class P of physicalistically acceptable predicates was supposed to have a structure. We supposed that P contained predicates

---

* I now believe that this answer is inadequate. In the Appendix of this book, a better answer is given.

from physics and inorganic chemistry and predicates which could be construced from these using first order operations. Physicalism was then defined as the doctrine that the language of the final theory contained only predicates from P. If we allowed in the final theory predicates other than these, we would defeat this "linguistic" version of physicalism: we would have to admit predicates not definable in terms of P. It might be that such predicates could be counted as physicalistically acceptable, but we would need to change our ground to some other as yet undefined notion of "physicalistically acceptable". For example, it might be that we should have to fall back on some primitive notion of "physicalistically acceptable property", and define "physicalistically acceptable predicate" in terms of that. These alternatives seem to me to be full of difficulties, and it would be better to avoid them.

This is all very well, but can there be non-synonymous predicates which express the same property? Many philosophers have thought so, and in my view they are right.

Consider first the question of whether there can be contingently true property identity statements. It is not difficult to show that if there can be true property identities, there can be contingently true ones. For there to be property identity statements, we need to have singular terms referring to properties. Let "a" be such a term. Then "a=a" is a true property identity statement, and "a=(ix)(x=a & p)" is true just when "p" is true.

Examples of property identities such as "Red is the colour of pillar boxes" can easily be seen not to be mysterious also. If we allow that "a" refer to a property,

then we allow that formulae of the form "Fa" can be true. (We allow that properties can have properties, if you like.) Some of these "second degree" properties might be relational properties of the original properties e.g. the relation between the property and those things which have the property. A property of properties such as the relation of having can presumably be had contingently by the properties which have it. So "Fa" may be contingent. Thus "a=(ix)Fx" might well be contingently true. "Red is the colour of pillar boxes" is evidently of this form.

Again, we might know that a certain property, a, is that property the possession of which by a given object is causally responsible for a certain event's occurring. If any such identity is true, then there are corresponding identities which are contingently true. Again, if "Fa" is ever true, where "a" refers to a property, then "F((ix)(x=a & p))" for some suitable "p", is contingently true, and so, as above, "a=(ix)(F((iy)(x=y & p))" is contingent.

These examples give some reason for thinking that nonsynonymous predicates can express the same property. Suppose that "a=b" is a contingently true property identity statement. Then "a" is not synonymous (in the language in which "a" and "b" are referring expressions) with "b". It seems reasonable to conclude from this that the predicates "x has a" and "x has b" are nonsynonymous. Now there might be some doubt about whether "x has a" and "x has b" express the properties a and b respectively. It seems to me that they do. (If "x is red" and "x is pillar-box-coloured" express the same property, then it seems reasonable to say that "x has the colour red" and "x has the colour of pillar boxes" express that same property.) But even if

they do not, at least they express what might be called particular instances of the having relation (which holds between things and the properties they have). But if a and b are the same property, then "x has a" and "x has b" would seems to express the same instance (which is a property) of the having relation, and thus the non-synonymous predicates "x has a" and "x has b" express the same property.

It might be objected to this last point that if a predicate , say "Fx", expresses the property a, then the predicate "x has a" cannot express the same property, for "Fx" and "x has a" have different ontological commitments: from the latter, but not the former, we can deduce "(Ex)(x has y)". Even if this is right, it does not defeat the argument that there is some common property which "x has a" and "x has b" express. But if a property - theory of predication is true, so that necessarily, if "Fb" is true, then there is some property y such that y is expressed by "Fx" and b has y, then "Fb" carries a commitment to properties, even if that commitment is not manifest in a language in which there are predicates but no quantification over properties. In a language in which there is quantification over properties, and in which the commitment of predicates to universals is explicit, we might easily have, as a theorem, "(x)(Fx → x has a)", from which we can deduce the ontological commitment of "Fx" to "(Ey)(x has y)".

These are complex matters, and I will not pursue them further. Gary Malinas[1] contrasts examples like the above property identity statements, with examples like the

---

[1] Malinas "Physical Properties". See bibliography.

well known alleged identity of temperatures and kinetic energies, which he calls "non-trivial", "non-contrived". Malinas, following Hilary Putnam,[2] offers a set of sufficient conditions for nonsynonymous predicates to express the same properties. In the next section, I will digress from the main argument, and try to give a better account than Malinas'.

2. Conditions for the Contingent Identification of Properties Via the Reduction of Predicates by Biconditionals.

Theories which do not contain names for properties present us with a problem. Under some circumstances, it seems right to say that a reduction by biconditionals gives grounds for holding that the reduced predicates are eliminable. Under other circumstances, it does not seem right to say this.[3] The problem is to say which circumstances are which. It is an important problem for the eliminative materialist who would wish to claim that because physiology will some day reduce psychology (by means of biconditionals), psychological predicates are eliminable. Such an eliminationist would have to discover just which conditions are attendant on the reduction, and show that those conditions are sufficient for elimination.

The eliminationist's problem connects with another. The conditions under which a predicate will be eliminable on reduction by biconditionals will be conditions under which the properties expressed by the eliminable predicate and by the predicate it is reduced to are identical. The whole problem might therefore be expressed as a

---

[2] Putnam 1970.

[3] For an example, see Causey 1972 p.414.

problem about the identification of properties: under what conditions do the predicates of a theory which reduces another theory by biconditionals express the same properties as those expressed by the predicates of the reduced theory?

Solving this problem will not solve the more general problem of when an arbitrary pair of (non-synonymous) coextensive predicates express the same property. I do not know how to solve this problem, and I will not attempt it in this book. In fact, I suspect that it will not have a very interesting solution, if only because predicates have a vast variety of different uses in different contexts. It seems to me that the best that can be done is to give necessary and/or sufficient conditions in particular contexts, for example, the context of reduction. Certainly, the context of reduction is particularly important for our purposes.

An example of sufficient conditions for nonsynonymous predicates to express the same property.

The following account of Malinas' views modifies them slightly but not significantly, I believe. Malinas thinks of the terms (predicates and individual constants) of a theory as divided into two classes, as others have, but his classes are respectively the class of terms which describe antecedently agreed on phenomena, and the class of terms used for explaining the phenomena, which, for a theory T, he calls Voc(T). The condition for contingent identification works only for predicates from Voc(T). Say that T_1 reduces$_1$ T_2, if T_1 reduces T_2 by means of biconditionals, $Voc(T_1) \subset Voc(T_2)$, and the biconditionals for all members of $Voc(T_2)$ are well established.[4]

[4]Kemeny and Oppenheim's term. See Kemeny and

$\phi \in Voc(T_1)$ and $\psi \in Voc(T_2)$, where ϕ , ψ contain just one free variable "x", are nomologically equivalent, if $\psi \notin Voc(T_1)$, T_1 reduces$_1$ T_2, and "$(x)(\phi \equiv \psi)$" is lawlike and true. Finally, a sufficient condition for two predicates ϕ , ψ to denote the same property, is that they be nomologically equivalent, or both be nomologically equivalent to a third.

Some brief comments: the condition that the biconditionals be well established seems unnecessary. It is an epistemological condition, and it is surely not essential that we be in a position to know or rationally believe that two predicates express the same property in order for them to do so. The lack of a condition that T_1 and T_2 be true also seems to be a defect. Voc(T) is also problematic, in that it relies on a distinction between explanation and description. Malinas also describes the descriptive terms as " 'old' terms relative to the development of the theory", and Voc(T) as the " 'new' terms introduced by a theory to explain what is described".[5] He says that this is the "kind of distinction" made by Lewis, and "old" and "new" certainly suggest this. But as we have interpreted Lewis, the "new" terms are those that get their meaning from the place they occupy in the theory. It may, of course, be that the two notions are coextensive, (perhaps via the idea of newness). On the other hand, there does not seem to be any reason why a term should not be an O-term in Lewis' sense, and yet not function in explanations of phenomena.

Oppenheim 1956.

[5] Malinas ibid. p.6.

Malinas' theory is included only as an example. I will proceed to offer a set of sufficient conditions for property identity which avoid the difficulties mentioned in connection with Malinas' view.

We consider the case where T_1 and T_2 are extensional and first order, and T_1 reduces T_2 by means of just one biconditional, "$(x)(F_1x \equiv F_2x)$", where "F_1x" ϵ $L(T_1)$, "F_2x" ϵ $L(T_2)$. The only other requirement we make is that T_1 is true. That the biconditional is true follows from the assumption of reduction. Modify T_1 and T_2 as follows: wherever "F_1" occurs in T_1, replace it by "f_1 ins", where "f_1" is an individual constant, and "x ins y" a binary relation. "f_1" is intended to be a name for the property expressed by "F_1x", and "x ins y" the instantiation (or having) relation. Call the resulting theory "T_1'". Similarly, replace "F_2" everywhere in T_2 by "f_2 ins", and call the result "T_2'". Now, on the assumptions that f_1, f_2 are properties instantiated in things just in case "F_1x", "F_2x" respectively are true of those things, we have that T_1', T_2' are true, and that T_1' reduces T_2' by means of the true biconditional "$(x)(f_1$ ins $x \equiv f_2$ ins $x)$". Let "$T_1'(x)$", "$T_2'(x)$", be the results of substituting "x" for "f_1" and "f_2" respectively throughout T_1' and T_2' respectively. Then we can substitute singular terms for "x" throughout $T_1'(x)$ and $T_2'(x)$. Thus, for example, $T_1'(f_1) = T_1'$. By an obvious extension of usage, we can speak of $T_1'(x)$, $T_2'(x)$ being <u>true of</u> things.

Now we cannot conclude, even in the conditions envisaged by Malinas, that $T_2'(x)$ is true of exactly one thing, f_2. Because T_2 is extensional, more than one predicate can stand in the place of "F_2x" in T_2, while

T_2 remains true. Any predicate coextensive with "F_2x" will do, e.g. "F_2x v x is a unicorn". Similarly, $T_2'(x)$ is (generally) true of more than one property if it is true of f_2. If disjunctive properties are permissible, as will be argued in Chapters Seven and Eight, and if disjunctive predicates can express disjunctive properties, then if "F_2x" is replaceable <u>salva veritate</u> in T_2 by the disjunct "F_2x v Fx", and if "F_2x v Fx" expresses the disjunctive property f, where $f \neq f_2$, then $T_2'(f)$ and $T_2'(f_2)$ are true. Alternatively, if conjunctive properties are permissible, then if "Fx" is coextensive with "F_2x", and "F_2x & Fx" expresses the conjunctive property, the property of being f_2 and f, where the property of being f_2 and f is not identical with f_2, then $T_2'(f)$ and $T_2'(f_2)$ are true. More importantly, if f_1 and f_2 are distinct properties, then since both $T_2'(f_1)$ and $T_2'(f_2)$ $(=T_2')$ are true, $T_2'(x)$ is true of two distinct properties.[6]

So we cannot make Lewis' assumption that the Ramsey Sentence for T_2' will have a unique realisation. If we can find conditions in which T_2' has a unique realisation, then we will have solved our problem, however, for T_2' has a unique realisation only if $f_1 = f_2$. (It is not unreasonable to make the assumption that T_2 has a unique realisation, however, for the sorts of reasons that Lewis gives.)

---

[6] This follows from the metatheorem, easily proved, that if T_1 reduces T_2 by means of the single biconditional "$(x)(Fx \equiv Gx)$", where "Fx" ε $L(T_1)$, "Gx" ε $L(T_2)$, then for any sentence "... G ..." belonging to T_2 and containing the predicate "G x", there is a sentence "... F ..." belonging to T_1, such that "... F ..." is the result of replacing "G" wherever it occurs in "... G ..." by "F".

We can, it seems to me, make the assumption that T_2' has a unique realisation, of a certain sort. If we suppose that T_2' has a unique realisation, then the problem of finding a suitable unique realisation for T_2' reduces to the problem of finding a single property of a certain sort for $T_2'(x)$ to be true of. Now considerations similar to those given by Lewis suggest that it is a methodological presupposition of scientific theorising that scientists are attempting to capture those properties the possession of which is causally relevant to the behaviour of objects in the domain of their theories. Classical thermodynamics, for example, was an attempt to describe the causal function of temperature properties and not properties like the disjunctive property of being 273 degrees Absolute or a unicorn, even though temperature predicates are true of things only if those predicates disjoined with "x is a unicorn" are.

"Causal relevance" is not an easy notion to spell out, and I will avoid doing so here. There is an excellent discussion of the notion in Peter Achinstein's paper "The Identity of Properties".[7] For my purposes, two considerations are relevant. The first is that "causal relevance" rather than "cause" was used, because there might be more than one property described by the theory and operating on a given occasion to cause the behaviour of objects which possess them. The second consideration is that it seems to me that it is part of the intuitive notion of causal relevance in the context of scientific theories that not any artificial construction of properties by conjunctions and disjunctions counts as

---

[7] Achinstein 1974.

causally relevant to the behaviour of objects which possess it, even if one or more of the "atoms" of the construction are causally operative on those objects. The property of being 273° A. or a unicorn, is not causally operative in thermodynamic systems obeying the laws of thermodynamics; only the property being 273°A. is.

If this is right, then we can say that it is a methodological presupposition of science that a theory like T_2 above is such that the reduced predicate "F_2x" expresses just one property (we have called it "f_2"), such that f_2 is the unique property x such that $T_2'(x)$ and the possession of x by objects in the domain of T_2 is causally relevant to those objects which possess it obeying the laws of T_2. ("Laws", as usual in this essay, means "sentences".) In symbols

(1) f_2 = $(ix)(T_2'(x)$ & the possession of x by objects in the domain of T_2 is causally relevant to those objects which possess it obeying the laws of $T_2)$

Now turn to f_1. If T_1 reduces T_2 by means of the single biconditional, and the biconditional is true, then f_2 is at least coextensive with f_1 (which is, of course, a necessary condition for property identification). Under these conditions, there are a number of alternative possibilities, only one of which is sufficient for the identification of f_1 with f_2. The first possibility is that an object's possessing f_1 is causally relevant for the object's possessing f_2. It seems clear that an object's possessing a property on a given occasion cannot be causally relevant for the object's possessing that property on that occasion. That is, f_1 cannot be identical with f_2. However, if " causally relevant to " is transitive, then the two properties, f_1 and f_2, will be such that $T_2'(x)$ is true of them and the possession of them is causally

relevant to ... etc. That is, in those circumstances, (1) is false. If (1) is true, then, we must rule out this possibility.

The second possibility is that while f_1 and f_2 are coextensive, f_1 is not causally operative in the behaviour of the objects in the domain of T_2. In this circumstance, if (1) is true, then $f_1 \neq f_2$. However, notice that if f_1 is not causally relevant to those objects which possess it and which are in the domain of T_2 behaving in the way that T_2 says they do, then nor is f_1 causally relevant to those objects which possess it and which are in the domain of T_1 behaving in the way that the closure of the union of T_1 and the biconditional says they do. This is because, since the biconditional is true, the domain of T_2 is a subset of the domain of T_1, and the behaviour which T_2 describes is included in the behaviour which the theory determined by the union of T_1 and the biconditional describes (T_2 is a subset of the closure of the union of T_1 and the biconditional). This is not to say that the possession of f_1 might not be causally relevant to the behaviour described by T_1 above. f_1 might be causally operative over the range of phenomena described by T_1, but have nothing to do with the possession of f_2 by those same objects in the domain of T_1: f_2 and f_1 are coextensive, perhaps even linked in a lawlike way, but the possession of f_1 is not causally relevant to the possession of f_2, and hence f_1 is not causally operative in that behaviour of objects described by T_2.

It is these two sorts of circumstances, I suggest, which typically obtain when we have a case of reduction by biconditionals without property identification. Either the possession of the reduced property is caused by the

possession of the reducing property, or the reducing property does not causally operate in the behaviour of the objects which the reduced theory describes, and the reduced property does so causally operate. I do not claim that these circumstances constitute necessary conditions for the non-identity of properties involved in reduction by biconditionals. A reason why it is difficult to give jointly necessary and sufficient conditions in these circumstances will be given below.

The third and last possibility before us, is that (1) be true and that the possession of f_1 be causally responsible for the behaviour of those objects which possess it described not merely by T_1, but by the closure of the union of T_1 and the biconditional. For then, since T_2 is a subset of that theory, the possession of f_1 is causally responsible for those objects which possess it behaving in the way T_2 says they do. In symbols

(2) $T_2'(f_1)$ & the possession of f_1 by objects in the domain of T_2 is causally relevant to those objects which possess it obeying the laws of T_2.

But from (1) and (2), we can conclude

(3) $f_1 = f_2$.

(1) and (2), therefore, are sufficient conditions for the identification of the properties expressed by "F_1x" and "F_2x" in the original theories T_1 and T_2.

Notice this: we argued that it was a methodological assumption that such theories attempted to capture just one property expressed by the predicate which was causally operative in the relevant way discussed above. The assumption guaranteed (1). When the assumption is false, then at least two coextensive properties are causally operative in the production of the behaviour of

the relevant objects. The assumption did not guarantee (2) however. The methodological assumption only guarantees of f_1 that it is the unique property the possession of which by objects in the domain of T_1 is causally relevant to the objects possessing it behaving in the way that T_1 describes; and that need not be the way that T_2 describes. It is when the biconditional expresses a property identification that we can take this further step. To put it slightly differently, we can make the property identification when we can say that the property f_1 is responsible for the behaviour of objects that T_2 describes.

Something slightly stronger is true. Suppose that T_1 reduces T_2 by means of "$(x)(F_1x \equiv F_2x)$", that T_1 is true, and that (1) is true. Then if (2) is true, $f_1 = f_2$. But also, if $f_1 = f_2$, it seems reasonable to conclude that (2) is true. (This might be questioned. Achinstein argues for a referentially transparent sense of "causal relevance" in which it is the case.[8]) Thus, if (1) is true, then a necessary and sufficient condition for $f_1 = f_2$, is (2).

Let me summarise the sufficient conditions given. We suppose that T_1 reduces T_2 by means of the single biconditional "$(x)(F_1x \equiv F_2x)$", where "F_1x" ε $L(T_1)$, "F_2x" ε $L(T_2)$, and that T_1 and the biconditional (and hence T_2) are true. Then a sufficient condition for the property expressed by "F_2x", f_2, to be identical with the property expressed by "F_1x", f_1, is that f_2 is the unique property the possession of which by objects in the domain of T_2 is causally relevant to those objects behaving in the

---

[8] Achinstein 1974.

way that T_2 describes, and f_1 is a property the possession of which by objects in the domain of T_2 is causally relevant to those objects behaving in the way that T_2 describes.

An obvious advantage of these conditions is that they avoid any mention of the idea that the biconditional be lawlike, unlike Malinas' conditions. There are several bonuses in this. One is that the notion of law-likeness is notoriously difficult to capture. Another is that it has been convincingly argued by both Causey and Achinstein that a lawlike and true correlation between predicates is not sufficient for property identification.[9]

Another advantage is that the sufficient conditions make no requirement that the predicates be members of Malinas' $Voc(T_1)$, $Voc(T_2)$. This seems right. I can think of no reason for supposing that only explanatory or theoretical terms can figure in property identifications.

There is a problem about making the conditions necessary as well as sufficient. The problem is that (1) might fail and yet $f_1 = f_2$. This circumstance could arise if T_2, while true, was <u>incomplete</u> in that another property, coextensive with f_2, was also causally responsible for the relevant behaviour of the objects which had them, but T_2 did not capture that property. In this circumstance, we might well have that T_1 also was incomplete in that it did not capture both properties, but it did capture the one which was identical with f_1. The sort of case I have in mind is where there is a

---

[9] Causey 1972, Achinstein 1974.

"hidden" property coextensive with f_2 and such that both co-operate in producing the behaviour of objects that have them, but as far as it looks from the standpoint of T_2, just f_2 does the work. I am unsure as to how to obtain necessary conditions, and I will not pursue the question here.

The case considered has been a particularly simple one, namely reduction by means of just one (monadic) biconditional. In the more general case where there are many biconditionals, I think that the problem can be solved too.

The following is a sketch of how I believe the solution goes. Suppose that there are n predicates and relations to be reduced in T_2, "$F_1^{k1}\ x_1, \ldots\ x_{k1}$", ... "$F_n^{kn}\ x_1, \ldots\ x_{kn}$", and they are reduced by biconditionals to the n predicates and relations of T_1, "$G_1^{k1}\ x_1, \ldots\ x_{k1}$", ..., "$G_n^{kn}\ x_1, \ldots\ x_{kn}$". The superscript in each case denotes the adicity of the relation. Substitute for each "F_i^{ki}" wherever it occurs in T_2 the expression "f_i^{ki} ins", where "f_i^{ki}" is a name for the property or relation expressed by "$F_i^{ki}\ x_1 \ldots\ x_{ki}$", and "ins" expresses the having relation. We obtain the theory T_2'. $T_2'(y_1, \ldots, y_n)$, analogously with before, is what is obtained by replacing each "f_i^{ki}" by "y_i", assuming, as we may do, that the "y_i" do not occur elsewhere in T_2'. Similarly, replace each "G_i^{ki}" by "g_i^{ki} ins" to give T_1'. Then T_1' reduces T_2' by means of n biconditionals, of which a typical member is "$(x_1) \ldots (x_{ki})(f_i^{ki}\text{ins}\ x_1 \ldots x_{ki} \equiv g_i^{ki}\text{ins}\ x_1 \ldots x_{ki})$". $T_1'(y_1, \ldots, y_n)$ is obtained analogously to $T_2'(y_1, \ldots, y_n)$. Then we can say that the n-tuple $(f_1^{k1}, \ldots, f_n^{kn})$ satisfies $T_2(y_1, \ldots, y_n)$, and the n-tuple $(g_1^{k1}, \ldots, g_n^{kn})$

satisfies $T_1'(y_1,\ldots,y_n)$. The more general sufficient conditions are (1) that $(f_1^{k1},\ldots,f_n^{kn})$ is the unique n-tuple of properties $(y_1,\ldots,y_n)$ such that $T_2'(y_1,\ldots,y_n)$ and the possession of one or more of $y_1,\ldots,y_n$ by those objects which possess them is causally relevant to those objects obeying the laws of T_2; and (2) that $T_2'(g_1^{k1},\ldots,g_n^{kn})$, and the possession of one or more of $g_1^{k1},\ldots,g_n^{kn}$ by those objects which possess them is causally relevant to those objects obeying the laws of T_2.

This completes the account of sufficient conditions for property identification that I will give. It is clear from the discussion that nothing that has been said prevents the biconditionals by which T_1 reduces T_2 from being contingent. Furthermore, it is clear that nothing that has been said requires that if "Fx" is reduced by biconditionals (so that "Fx" ϵ $L(T_2)$), "Fx" must also ϵ $L(T_1)$. If the biconditional by which "Fx" is contingent and "Fx" ϵ $L(T_1)$, then if the reduction satisfies the conditions for property identification, we have conditions sufficient for the elimination of "Fx" from theories which deal adequately with the areas covered by T_1 and T_2. If the reducing theory is the final theory, then "Fx" is subject to the sort of eliminability which the version of eliminative materialism discussed in this chapter needs.

3. An Argument For Physicalism.

If the contingent identification of properties of the sort described in the previous section is possible, then it would seem that we have a quick argument for physicalism. We can ignore the problem of analysing mental predicates to show that they are members of P,

and we can ignore the difficulties of denying that the mental predicates are ever true of anything. Logic allows that mental properties can be contingently identified with physical properties, and the likely future course of science makes it reasonable to believe that the identification can be made.

This argument has considerable persuasive force, but it suffers from a defect which has also emerged in earlier discussions of other physicalist positions. The defect is that the reason for the belief that the likely future course of science will be favourable to physicalism is only that ~~provided~~ if no good argument can be given to the contrary, we ought to believe in physicalism. But there might be such an argument, and I will claim that there is one. The place to look for such an argument is in the nature of mental properties, and the place to look for information about the nature of mental properties is in what, if anything, we introspectively know about them. It might turn out that we know too much about mental properties for them to be identified with physical properties. It might turn out that we know too much about our minds to be able to deny with Rorty, that mental predicates are true of things. It might turn out that we know too much about our minds to be able to give an adverbial semantics for mental predicates. It might turn out that some of the facts we know about ourselves cannot be given a physicalist or topic neutral analysis.

This supplies the motive for the investigation in Part Two, but we are not yet in a position to undertake it. In the next chapter, we will discuss a version of physicalism due principally to Hilary Putnam which claims to be albe to avoid the Identity Theory. The theoretical tools which have been developed so far will help us to see its strengths and weaknesses.

CHAPTER SEVEN. FUNCTIONALISM

1. Various Functionalist Theories.

In this chapter we will examine a version of physicalism which denies the Identity Theory. It has been called "functionalism", and its principal proponents have been Hilary Putnam and ~~Jeremy~~ Jerry Fodor. The discussion derives partly from an excellent article by Lycan,[1] who himself advocates a form of functionalism. It will be argued that functionalists do not succeed in establishing their case against IT.

Functionalism is the conjunction of two claims: first, that pain is a functional state, and second, that functional states are not neurophysiological states. The conjunction of the two claims is apparently inconsistent with the Identity Theory, though, as we will see, this is something of a moot point, depending on how "neurophysiological state" is construed. Certainly functionalism's proponents intend both that it be inconsistent with the Identity Theory and that it be consistent with materialism. We distinguish three main accounts of what a functional state is.

Fodor. "To say that the psychologist is seeking functional characterisations of psychological constructs is at least to say that ... the criteria employed for individuating such constructs are based primarily upon hypotheses about the role they play in the etiology of behaviour."[2] "... the hypothesized psychological constructs

---

[1] Lycan 1974.

[2] Fodor 1968 p.140.

are individuated primarily or solely by reference to their alleged causal consequences." [3]

Putnam. Functional states are logical states of some machine. We should not take "machine" too literally here. Putnam uses an analogy with Turing machines. An adequate definition of a Turing machine is as follows. A Turing machine has an input tape on which is printed a sequence of symbols from a finite alphabet, a reading head that scans one symbol at a time, and a finite set of internal states that it can be in. For a given initial internal state and a given symbol scanned, the machine can do one of four things. It can replace the symbol scanned by another symbol (which may be the same) and go into another (which may be identical) internal state, it can move left one symbol on the tape and go into another internal state, it can move right one symbol and go into another state, or it can halt. If the alphabet of symbols for a particular machine is the set $S = \{ S_1, \ldots, S_n \}$, the set Q of possible internal states = $\{ q_1, \ldots, q_m \}$, and "R", "L" represent the acts of acts of going right one square and going left one square respectively, the Turing machine may thus be represented by a set of quadruples (a_i, S_j, X, q_k) , where $q_i \varepsilon Q$ is the initial state, $S_j \varepsilon S$ is the symbol scanned, $X \varepsilon \{R, L\} \cup S$ represents the acts of going right, of going left, or of printing some symbol from S, and $q_k \varepsilon Q$ is the final state. The set of quadruples is called the "machine table".

---

[3] Fodor 1968 p.141. In an earlier article, (Fodor 1965), Fodor appears to contradict this. (See e.g. p.233.) I will not explore this difficulty for Fodor.

As is well known, it can be shown that a large number of apparently more complex machines can have their job done by some Turing machine. It is a conjecture of Church that for any machine that can effectively perform some computation, there is a Turing machine which can perform the same computation in a finite number of steps. It remains a conjecture because of the vagueness in our intuitive notion of "effectively", but no clear counterexamples have yet been found.

In claiming that pains are logical states, Putnam allows that the human machine might not be deterministic in the way that a Turing machine is: the transition from state to state might occur with varying probabilities.[4] This would complicate the machine table by requiring the introduction of a suitable matrix of probabilities. The theory of machines, the internal transition between states of which occurs with varying probabilities, has been extensively investigated.[5] It is an unnecessary complication for our purposes, so we will remain with Turing machines.

For Putnam, q_i are "logical" states in that a machine table does not contain any description of the make-up of these states, only how they occur in the overall working of the machine.

> ...the "logical description" (machine table) of a Turing machine does not include any specification of the physical nature of these "states" - or indeed, of the physical nature of the whole machine. (Shall it consist of electronic relays, of cardboard, of human clerks sitting at desks, or what?) In other

---

[4] Putnam 1967 p.155.

[5] e.g. Ashby 1964.

> words, a given "Turing machine" is an abstract machine which may be physically realised in an almost infinite number of different ways.[6]

Lycan. In order to describe Lycan's position, we must distinguish between two ways of understanding the predicate "x is a Turing machine". One way, somewhat characteristic of mathematicians, is to identify the Turing machine with its machine table. A Turing machine is thus a set of ordered n-tuples (and the elements of the n-tuples need be nothing physical) and so is some sort of platonic object. For example, Martin Davis writes

> Definition 1.3 A Turing machine is a finite (nonempty) set of quadruples that contains no two quadruples whose first two symbols are the same.
>
> The q_i's and the S_i's that occur in the quadruples of a Turing machine are called the internal configurations and its alphabet, respectively.[7]

Putnam speaks similarly (above) of a Turing machine's being "abstract"

The other way is to speak of actual physical machines as being Turing machines, in the way that one might say that a digital computer is a Turing machine but an analogue computer not. Some definitions might help the discussion. An ATM (abstract Turing machine) is a set of quadruples as specified above. The relation "is a realisation of" will be taken as primitive, but some explanation is in order. What makes a physical object a Turing machine can be expressed by saying that there is a correlation, some sort of paralleling, between the states of the object and states of the Turing machine.

---

[6] Putnam 1960 p.147.

[7] Davis 1958 p.5.

(ATM). Defining "is a realisation of" might not be such a difficult matter if we only needed to associate the states of the object with the q_i of an ATM, but there must also be some sort of physical (if the object is physical) operation corresponding to the input of symbols on the tape, and there is no precise way of saying what sorts of physical things count as being symbols to be read, or the act of scanning. Similarly also for the acts of moving left and right. Deciding that something is a realisation of an ATM is more or less just a matter of our being able to see that a set of physical processes somehow mirrors an ATM.

We place one restriction on the realisation relation: that something which is a realisation of an ATM be a temporal device. When in operation the realisation is sometimes in some states and (typically) at other times in others. This restriction is intended to prevent selecting any 4n objects and forming them into n quadruples to be a realisation of an ATM with n elements. Notice also that the realisation relation is a many-many relation. One Turing machine (ATM) can be realised by more than one object, and vice versa. We can now define a PTM (physical[8] Turing machine) thus:

"x is a PTM" $=_{df}$ "(Ey)(y is an ATM & x is a realisation of y)".

We should not assume that a PTM's being in a certain state (one of the states corresponding to the states of the ATM) and "scanning" a "tape" brings about the change of state, etc., for the changes of the states, and

---

[8] Notice that a PTM is not necessarily literally physical. The correspondence could presumably hold between an ATM and a ghostly PTM.

the states themselves, might all be brought about by some underlying process, rather than bringing one another about. This does not imply that the states of a PTM must be causally inefficaceous, only that they need not be causally efficaceous on one another. Also, it does not deny that in such a circumstance it might be reasonable to conclude that the underlying processes bringing one another about could also be thought of as a Turing machine. States of a PTM might be causally relevant to one another, but need not be.

Some states of a PTM might not be among those which correspond to the states of a Turing machine, others will be. We can call the ones that do, functional states. There will be other features of a PTM relevant to its being a PTM, viz those features corresponding to the tape, to scanning, etc., which are presumably not states. We can call these, together with the functional states, functional elements. (Then, by a circular but perhaps illuminating move, we might define the realisation relation by saying that x is a realisation of y if there is a 1-1 function between functional elements of x and elements of the quadruples of y.

We can now state Lycan's functionalism as follows: mental states are functional states of some PTM.

All these versions of functionalism are similar to Lewis' theory in an important respect. They share the idea that whether something is a mental state of a certain sort is determined by the connections it has, not just with stimuli and responses, but with other mental states (actually, stimuli and responses represent a difficulty for functionalism, which we will return to later). Any identification which we might try to make of functional

states with neurological states could not be a "state-by-state" identification: we should have to identify all the states taken together with their neurological counterparts.

There are at least two, and perhaps three, differences between functionalism and Lewis' theory. The first difference is that the identification of pain with a functional state is not obviously intended to be an analytical identification. Putnam, and Lycan following him, hold that the identification of pain with a functional state is made on "highly theoretical, but not conceptual grounds".[9] The second difference for Putnam and Lycan is that the picture presented is not a causal picture, but a functional one. This difference in PTM's, is that functional elements of a PTM are identified by their correspondence with (abstract) parts of an ATM. This is not so informative; it is less accurate but perhaps not too far wrong to say that where causal pictures use descriptions like "this occurred, and it caused that to occur", logical descriptions of Turing machines are like "this state occurs, then that one occurs". The difference that there seems to be is that functionalist descriptions are not causal, but they are similar to Lewis' descriptions in that they are descriptions of a succession of states. For Fodor, on the other hand, functional descriptions are causal descriptions.

The third (possible) difference, is in what is being claimed by functionalists not to be identical with neurological states. For Putnam, pain and functional states are properties, but we will see later that the sense of "property" here is not clearly a sense in which an Identity theorist would claim that such properties are neurological states.

---

[9] e.g. Lycan 1974 p.49, p.57.

2. Putnam's Arguments For His Position.

I propose to restrict myself largely to discussing Putnam's position. The arguments given do bear on Fodor's views, as will be clear. An argument of Lycan's will be discussed briefly.

There are two important arguments, not always clearly distinguished, that Putnam gives for his conclusions.[10] The first argument is that a functional state cannot be identical with a neurological state, for a given Turing machine can be realised in various ways. If, that is, we identified the functional state, **pain**, with one neurological state, we ought to identify the same functional state with say, a state of the diodes of some computer which also was a realisation of the Turing machine. But then by the transitivity of identity, we should have to identify some neurological state with a state of diodes, and that is absurd. The second argument, endorsed by Lycan, is that pain cannot be identical with a neurological state, for then nothing but organisms with human or near-human physiology could have a pain.

(2.1) The first argument suffers from a confusion of ATMs and PTMs. (It is significant that Putnam does not make this distinction.) It is sound if it is an argument for the conclusion that states of ATMs cannot be neurological states, but then it can hardly support a thesis contrary to the Identity Theory, for it would be extremely peculiar to identify pain with a state of an ATM. Pains,

---

[10] Putnam gives other arguments, usefully summarised in Lycan 1974. See also Kalke 1969. We will not discuss them. Lycan's arguments against them seem conclusive.

if they are states at all, are states of humans, and if humans are any sort of Turing machines at all, they are PTMs. If, on the other hand, it is an argument for the conclusion that the functional states of PTMs cannot be neurological, it has a false premise. It is a mistake to think that a PTM can be "realised" in various ways. The PTM is a realisation.

To characterise a state of a PTM as functional is not to say anything about its being neurological. That does not prevent it from being neurological, however, for to characterise a state of a PTM as functional is to say something about the existence of a 1 - 1 function which, with the state as argument, takes some state of an ATM as value. Nothing prevents a neurological state from being an argument of such a function.

(2.2) Putnam's second argument is deceptively simple. Certainly no Identity theorist that I have heard of has wanted to say that beings with non-human physiologies logically could not have pain. (It is, of course, a matter of whether the Identity theorist is committed to this.)

In support of his argument, Putnam assimilates pains (or the state of pain) to properties.[11] (They are, according to him, functional properties.) And if the property, pain, were the same as some neurological property, then it could not also be the same as some property involving diodes. It follows from this that pain cannot be a neurological state (states are properties for Putnam).

A useful analogy is between functional properties and causal properties. Take a typical causal property:

---

[11] Putnam 1967 pp.150-155.

the property of being the cause of John's cry. It seems clear that we cannot identify this with a certain sort of neurological property, say the property of being a type A neuron.[12] Another, similar, analogy is with the spatial properties of the brain e.g. being 2cm. directly above the Rhinencephalon. No-one ought to be tempted to identify this property with the property of being a type A neuron, even if there were exactly one type A neuron in the brain and it were 2 cm. directly above the Rhinencephalon.

If pain is a functional state and functional states are functional properties, then why should we not agree that functional properties, like such causal and spatial properties, cannot be identical with properties like being a type A neuron? Otherwise, only things with type A neurons could have the right sort of functional (or causal or spatial) property, and this is surely false. An example of a functional property which intuitively would fit this argument, is the property of being a realisation of the state q_1 of the ATM, T.

Now there are very many versions of what the Identity Theory is supposed to be identifying with what: states with states, properties with properties, events with events, processes with processes. Distinguishing between these various categories is a murky business and

---

[12] It seems clear, but I am not sure how to prove it. One reason for believing it is that the first is somehow relational, the second not necessarily (type A might be determined by their mass). Another reason is similar to Putnam's: that type A neurons might not have been the cause of John's cry. There seems to be something right about this reason, even though something parallel could be said against any contingent identification of properties, including true ones.

one I hope to be able to avoid.

One clear claim of most Identity theorists, however, and certainly a claim of the Ramsey Sentence theorists, is that the sort of things that are the causes of behaviour and the effects of stimuli are identical with neurological entities. They do not so much want to identify causal properties of the brain such as being the cause of B with certain neurological properties, but what they are the causal properties of, with what they are the neurological properties of. This is not to deny that what they are the properties of may not also be properties. We should not wish to restrict our categories so much. (Evidently properties are causally involved in what occurs, even if we would not want to say that properties are causes.) In the particular case, the Identity theorist claims that whatever it is about pain that causes John's cry (a pain-state, a pain-event, a pain-property, it is not so unusual to call it "pain") is identical with the neurological entity.

Does this commit the Identity theorist to holding that pain can only be had by near-humans? Not at all. Fodor is just wrong here. The cause of John's cry is neurological, the cause of the Martian's avoidance behaviour is a state of its silicon crystals.

It might be thought that this much can be salvaged from Putnam's position. Perhaps pain is not a functional property, but nevertheless, to say that someone is in pain is to ascribe to them a certain functional organisation. That is to say, the property being in pain is a functional property and so cannot be identical with a neurological property.

The answer to this is twofold. The first part of the answer is that there are perhaps more neurological properties than Putnam imagines. Consider the analogy with causal properties again.

The discussion will be aided by distinguishing two ways in which a property might be said to be a "causal property". The first way is when the property is a causal relation, such as the relation: being the cause of x. The second way is when a property is said to be causally operative in the production of something. We might, as we did in Chapter Six, characterise a property as the property the possession of which is causally responsible for the production of x.

Now it seems to me that at least some properties associated with pains are causal properties of the second sort. If a person has the property, being in pain, and so groans, then the person is in a state or has a property the possession of which by the person is causally responsible for the groan. If this is right, then the discussion in Chapter Six should lead us to expect a contingent identification of being in pain with neurological properties such as the property of having one's synapses in the state of potential-distribution p, or the property of having type A neurons.

Putnam might reply that such a move could not account for the intuitively obvious fact that things with non-human physiologies can have pains, since it identifies being in pain with specifically human states. This forces us to make a modification to what was said in the previous paragraph. Smart has supplied strong grounds for allowing the modification, it seems to me.[13]

---

[13] Smart 1971, section IV.

Rather than identify the property being in pain with the property having one's synapses etc., we should have to identify it with the disjunctive property: having one's synapses in state S or having one's diodes in state S^1 or ... . Disjunctive properties, as Smart convincingly argues, are perfectly objective properties.

Far from finding this counterintuitive, I think that we must be prepared to accept it. Stand a human in pain and a Martian in pain side-by-side. Now if we want to say that being in pain is the property the possession of which is causally responsible for their behaviour, then we must conclude that the (single) property causally responsible for their behaviour is a disjunctive property. Disjunctive properties will be discussed further in the next chapter.

So if Putnam were to agree that functional properties were causal properties in the second sense, then we must conclude that he has not made out his case against the Identity Theory. But Putnam might be disinclined to believe that functional properties, and being in pain, are causally operative in human behaviour. He might, however, still wish to preserve the notion of causality in the notion of pain. Denying this would be unacceptable in my view. If Putnam wished to do this, then he could hold that being in pain is a causal property in the first sense. That is, he could hold that it is a relational property like being the cause of B . This would exempt him from identifying it with disjunctive properties. More importantly, it would have the advantage (for Putnam) that it would not require him to identify it with distinct properties like being in neuronal state S and being in silicon state S^1.

But if being in pain is such a causal property, why should it not be identical with a neurological property? Indeed, why should it not be a neurological property? Neurophysiology contains more predicates than "x is a type A neuron"; it also gives accounts of causes. "x is the cause of B" is a typical neurophysiological predicate. Surely no Identity theorist would deny that the relational property of being the cause of B is a perfectly acceptable neurophysiological property.

Finally, Putnam might wish to sever the concept of being in pain from causality altogether. As argued above, this seems quite wrong. But suppose he is right. Suppose that being in pain is to be assimilated to properties like the property of being a realisation of state q_i of Turing machine T (cf. the spatial properties of the brain). I think that this is closest to the spirit of what Putnam says. But even if he is right, then in order to establish that being in pain is not a neurophysiological property, he would have to establish that such properties are not typically alluded to by neurophysiologists when on the job. It is, of course, just not true that neurophysiologists do not investigate the "logic design" of the brain. This part of the reply to Putnam, then, is that even if Putnam were right in claiming that certain mental properties are non-causal functional properties, and even if such non-causal functional properties were not neurophysiological properties, they ought to be. Putnam would have to be thanked cordially for pointing to a new direction in research, but he has hardly established anything that an Identity theorist should be concerned to deny.

To sum up what has been so far said: Fodor's functionalism is met by pointing out that the causes of

behaviour can be identical with different states in different species, and that the Identity Theory does not deny it. Putnam's functionalism is met by pointing out that the core of the Identity Theory is the identification of causes and causally relevant properties with neurophysiological states and properties, but that in any case functional properties and causal properties of either kind are neurophysiological properties.

We mention one last argument, from Lycan. If pains were identical with neurophysiological states in humans, and with silicon states in Martians (i.e. if being in pain is a disjunctive property), then how could we give an account of why the word "pain" could properly be used of both of them? The answer is that Lycan is making an absolute-relative confusion.[14] If to say that things are in pain is to say that they have the property causally responsible for the production of B, then what the properties have in common is something relational. If to say that things are in pain is to say that they are in some state corresponding to a q_i of an ATM, then what the states have in common is a relation to an ATM. Lycan's argument only looks plausible if you are looking for something <u>intrinsic</u> to pain in common to all pains. It would not surprise me if Lycan thought this because he was thinking as functional states of PTMs as a bit like baby q_i's of ATMs (which was Putnam's mistake), i.e. as something intermediate between states of ATMs and the neurological states of PTMs; instead of seeing that to characterise a state as functional is to characterise it relationally.

---

[14] See my <u>Absolutes and Relatives</u>, forthcoming.

This completes my discussion of functionalism. I conclude that functionalism has not made out its case against the Identity Theory.

So far, I have explored various ways of reconciling the use of mental predicates with a physicalist world view. The ways have mostly clustered around the ideas of reduction and elimination. It has recently been claimed by Donald Davidson and Hilary Putnam that it might be unnecessary either to reduce or eliminate mental predicates, and yet physicalism be true. According to Davidson, mental predicates are not reducible to physical predicates, and laws governing the mental are not deducible from laws governing the physical. In the next chapter, the last in this part, we will examine this approach to physicalism.

CHAPTER EIGHT. IS THE MENTAL IRREDUCIBLE?

1. The Anomalousness of the Mental.

It might be thought that our definitions of physicalism, reduction, elimination and property identification impose too strict a methodology on the enterprise of reconciling the use of mental predicates with physicalism. To say that the mental could be reduced to the physical is to imply that laws containing mental predicates could be deduced from physical laws, perhaps with the aid of suitable bridging laws. At points in this essay I have spoken as if this is a possibility to be taken seriously. However, some recent authors, particularly Hilary Putnam (who has now repudiated functionalism)[1] and Donald Davidson,[2] have argued that physicalism could be true while there are no laws relating to mental-mental or mental-physical interactions and correlations. Davidson calls this latter claim the principle of the anomalousness of the mental. In this chapter, I will discuss whether this claim is true. I will not be concerned with the details of Putnam's position. In the next section I will discuss some parts of what Davidson says.

There seem to me to be at least three considerations which are inclined to show that there could not be any laws in psychology, and, correlatively, that psychology or "common sense psychology" could not be deduced from a theory (perhaps augmented by biconditionals) all

---

[1] Putnam, "Reductionism and the Nature of Intelligence", see bibliography.

[2] Davidson 1970, 1974.

of whose predicates come from physics or inorganic chemistry. These considerations certainly show that the old empiricist picture of the Unity of Science, wherein there is a reduction of sociology to physics via a series of interim reductions: sociology-psychology, psychology-biology, biology-chemistry, chemistry-physics, is mistaken. I will argue in a later section that these considerations do not force us to give up the methodology developed in Part One of this essay.

It is not uncommon to speak of sociology as being a "higher level" science than psychology, psychology being a higher level than biology, etc., and we will adopt this terminology. Now the first reason for thinking that the mental might be anomalous, is that at higher levels, there are propositions taken into account which do not come from any laws, i.e. lawlike propositions, of the lower level discipline. Relative to those lower laws they are at least accidental generalisations. For instance, there are no laws of chemistry which say that there must be cells; there are no laws of biology which entail that something with roughly the mental make-up of humans will ever come into existence. If generalisations about such entities are a necessary part of the higher level science, then not only can we not obtain the truths of the higher level science from the collection of lawlike propositions of the lower level science, but there is reason to think that at least some of the truths of the higher level science are not lawlike. However, even if we grant that lawlikeness distributes over entailment, this point does not go against what has been said earlier in this essay. Theories were conceived of as sets of sentences, not sets of laws. Certainly the theory of property identification given in Chapter Six did not turn

on there being lawlike connections between predicates which express the same property.

The second reason for thinking that the mental might be irreducible to the physical is as follows. Psychology employs e.g. the concept of a person's body and its parts; biology employs the concept of a cell. Neither of these concepts has a precise definition in terms of the entities which make it up. Cells, for instance, do not have an exact chemical composition. Yet "laws" in biology sometimes essentially mention cells, and similarly with bodies in psychology or common sense psychology. So not only will we not be able to obtain higher level laws from lower level laws by deduction, but also the generalisations on the higher level can never be better than loose, probabilistic generalisations.

The third reason is that explanations and descriptions in the higher level science often have to import concepts from a discipline at a lower level than the prima facie reducing discipline. An explanation of an event in a cell might turn, among other things, on the phenomena of electromagnetism, gravity and radioactive decay, which are at a deeper level than chemistry. A psychological event, e.g. a person's feeling pain, might need for its description and explanation facts about the transmission of heat in metals. It seems to me that this consideration shows that any reduction of higher level sciences must be reduction straight to the physical. Any interim reductions must at best be partial aids to the full reduction to physics.

One argument for the irreducibility of various sciences which I do not think is a good argument is as

follows. Somebody might argue that higher level sciences employ concepts which have no definition in the respective lower level science, on the grounds that any attempted definition would have to be circular. Consider the example of sociology. It might be argued that the concept of an army could not be defined without concepts like those of troops, command, etc. But those concepts could not be defined without the concept of an army. Thus any definition into the concepts of the lower level science (psychology) would leave something out.

But to say that an analytical definition of a single term like "army" could not be given in the concepts of (say) psychology does not prevent an analytical definition of some of the terms of a whole theory into a reducing theory (nor does it prevent a contingent identification of the properties expressed by those terms with properties expressed by terms in the reducing theory). The Ramsey Sentence approach can help us here. Treat the terms "army", "troops", "command" as all theoretical concepts in a theory within sociology, defined by the place they occupy in the theory, and then reduce all the terms together. Surely the situation is similar between sociology and "common sense psychology", in that concepts like that of an army derive their meaning from an interlocking set of concepts and relations between a variety of entities, theoretical and non-theoretical.

2. Davidson on the Anomalousness of the Mental.

Davidson has a different argument for the anomalousness of the mental. The mental for Davidson is characterised in a Brentano-Chisholm fashion,[3] as being essentially connected with a certain referential opacity

---

[3] Chisholm 1967, 1957.

and non-truth functionality possessed by verbs of propositional attitudes like belief, wanting, desiring, etc. There are considerable problems in defining the mental this way, as has been argued by Kim,[4] and as Davidson himself notes. We will not explore these problems, but even if the mental cannot be isolated in such a way, at least verbs of propositional attitude constitute a subclass of mental terms to which Davidson's argument applies, and if he is able to establish his conclusion for just these terms, it is a strong conclusion.

Briefly, Davidson's argument is this. We cannot separate the ascription of propositional attitudes to a person from the ascription of meanings to their utterances. To determine what a person believes involves deciphering their utterances, and translating their utterances involves knowing such things as what they mean by them and what their beliefs are on the occasion of utterance. The translation of utterances, however, involves the Quinean problem of the Indeterminacy of Translation.[5] Insofar, then, as the indeterminacy in correctly translating a person's language always leaves open the possibility of a radical revision in our translation manual for the person's language, any theory of what a person's beliefs, wants, etc., are must always be open to a similar radical revision. We can never rule out the possibility of having to make this revision in our theories about people in the light of further facts. Scientific theories about the physical, on the other hand, have a certain convergence property. As the theories develop, it becomes less and less likely

---

[4] Kim 1970-1.

[5] e.g. Quine 1960.

that we will have to make a major revision in them. Therefore, ascriptions of propositional attitudes are different in kind from the application of physical predicates. In particular, laws relating to the mental could not be deducible from wholly physical laws (for if they could then they would have no more indeterminacy than physical laws), and, as a corollary, there could not be true biconditionals linking verbs of propositional attitude with physical predicates.

I do not wish to discuss here the merits or otherwise of the principle of the Indeterminacy of Translation. I do wish to argue, however, that it is misapplied here. I will argue in the form of a dilemma.

Either Davidson holds a strong form (to be outlined shortly) of what I will call the Indeterminacy of the Mental, or he does not. Suppose he does. What is this strong form of the Indeterminacy of the Mental? I will take it as the denial that there is a single correct ascription of propositional attitudes to a person at a time: either there are no correct ascriptions of propositional attitudes, or there are many, equally correct. This interpretation of what Davidson says has two merits: it is plausible to ascribe to Quine an analogous interpretation of the Indeterminacy of Translation,[6] and it fits with Quine's comment on Davidson's position, that "belief is invented".[7]

Now it seems to me that the denial of a single correct ascription of propositional attitudes has a consequence that there are no truths about human beliefs

---

[6] see e.g. Putnam 1974.

[7] Quine 1974 p.325.

for belief-verbs to capture, for what could "correct" mean here but "true"? But this in turn would seem to imply that there are no determinate facts about human beliefs. But this would be very curious. It would be strange to say either that human beings had no beliefs, or that they had beliefs, but no determinate, i.e. particular, beliefs. This interpretation of Davidson has the virtue that it seems to be parallel to the Quinean version of Indeterminacy of Translation in that that principle goes with the denial of any facts about synonymy in which to ground the notion of the correctness or truth of a translation. But, if I am right, it has the consequence of the denial of any determinate beliefs on the part of human beings, and that is surely unacceptable.

On the other hand, Davidson might be willing to concede that there are correct ascriptions of beliefs, etc., to humans, but that the indeterminacy arises solely because of the always-present possibility of radical revision in our ascriptions of propositional attitudes. The "evidence" does not "tie down" our theories about humans enough ever to warrant confidence that we will never have to undertake a major revision of them.

But how would this make for a difference between the mental and the physical? First, is it not precisely the situation with high level physical theories, that a variety of theories are consistent with the "evidence" and that Principles of Method are needed to decide between them? Even if we hold that all observation is theory laden, we must surely continue to hold that Principles of Method are needed to decide between certain competitors. Indeed, if all observation were theory laden, then would

we not be in the position of having to say that all ascriptions of physical predicates have this sort of indeterminacy?

Second, to what extent are our theories which include belief predicates really open to radical revision? We are faced here with a real-life situation. When we first come to know people, we do start off with little or no knowledge of what their beliefs and attitudes are. Certainly, as Davidson points out, it is proper to start with some presuppositions about people, for instance that most of their beliefs are true, and that they speak more or less the same language we do. But it is noteworthy that when we come to know a close friend very well, ultimately we get a very good idea about a lot of his or her beliefs and attitudes and language. Agreed, we always have in some sense the possibility of a radical revision of our theory about their minds. However, it is also true that in those cases where we do have to make a major revision in our theory about our close friend, it is proper and rational to be very surprised at having to make this change. If this surprise is rational, then surely it is the case that it was rational of us to have come to the beliefs about the person's mind that we hitherto had. If Davidson thinks that there is a difference between ascriptions of propositional attitudes and the application of physical predicates in this respect, he would seem to be condemning us to a radical scepticism about ever coming to know about the minds of our friends.

I conclude that Davidson's argument for the anomalousness of the mental is mistaken. Davidson's "proof" of physicalism from the anomalousness of the

mental deserves noticing. Davidson assumes (1) that there are mental-mental and mental-physical causal relations, and (2) that for any singular causal statement, there is an underlying law relating the causally-related items under some description. He then argues: let m, a mental event, cause p, a physical event. Then, by (2), there is an underlying law which relates m under some description to p under some description. These descriptions cannot be mental descriptions, for it has been argued that there are no strict mental-mental or mental-physical laws. They must, therefore, be physical descriptions. That is to say, m has a physical description and so is a physical event.

This argument assumes that the only descriptions other than Brentano-style intentional descriptions under which events could be causally related, or related in a lawlike way, are physical descriptions. It supposes that if there are no mental-physical laws, then the only laws which relate the events in question are physical-physical laws. But to suppose this is to ignore the possibility that there could be laws relating the mental and the physical in which the mental is characterised in a different way from Davidson's way. Furthermore, if we agree that there could conceivably be alternative descriptions under which there are mental-physical laws, then we will have to agree that the reasons for preferring underlying physical-physical laws are Ockhamist reasons. If there could be properly dualist causal relations, then the reasons for not believing that there are any, are reasons which derive from Principles of Method. But, as I have said before, there might be arguments to defeat such reasons. It might be that we can show that the hope of a wholly physicalist description of mental

phenomena is mistaken. At least, the possibility of producing such an argument ought not to be ignored. Davidson's "proof" of physicalism, therefore, is inadequate.

We are still left, however, with the reasons given in the first section of this chapter for the irreducibility of the mental. In the next section, I will try to show that those arguments do not establish what they might seem to establish.

3. A Note About Disjunctive Properties.

The anomalousness of the mental might appear to force us to take an entirely different approach to the justification of physicalism. If we cannot deduce a person's common sense psychology from any theory of physics, then presumably we will have to eliminate the predicates, either by denying their instantiation or by contingently identifying the properties expressed by mental predicates with physical properties. If it is true that we have mental properties, beliefs, attitudes, then the first sort of elimination cannot be made. If the mental is anomalous, then it is plausible to think that there are no biconditionals linking mental predicates with physical ones. But if there are no true biconditionals, how could there be contingent identifications, for it is surely a necessary condition for two properties being identical, that they are had by the same things.

This problem might lead one to change the definition of physicalism we have given in order to preserve physicalism. I cannot see any clear way of doing this, however, unless we fall back on some primitive notion of a physical property, and then claim

that mental properties are physical properties. But the difficulty in doing this would be to give some non-arbitrary reason for thinking that mental properties <u>are</u> physical if they are not coextensive with any properties dealt with explicitly in physics. For this reason, I think that a would-be physicalist should be loath to give up having some such definition of physicalism as the one we have given. At the same time, it would be surprising, it seems to me, if we could establish the falsity of physicalism by the sort of considerations we have given.

My reason for thinking this is that the problems of irreducibility that we have indicated are by no means confined to common sense psychology or common sense psychology augmented with a translation manual for the person's ideolect. Biology would also seem to have problems about irreducibility. Consider again the concept of a cell. The variability of chemical composition of a cell ensures that no definition of "x is a cell" in terms of predicates from chemistry describing the composition of a cell is possible. That same variability gives reason for thinking that at least some laws of biology essentially using "x is a cell" would be at best probabilistic. Finally, the fact that some interactions involving cells involve specifically physical, rather than chemical phenomena (e.g. the effect of radioactive material on cell composition) shows that any "reduction" of biology would have to be directly to physics, rather than via an intermediate step of "reduction" to chemistry. Now it would seem unreasonable to deny that there are cells, and hence if physicalism as we have defined it is to be saved, the property, being a cell, would have to be identified with some physical property, otherwise physicalism would have to be abandoned. And surely the existence of biology

does not refute physicalism!

I will discuss this problem only in connection with psychology. In my view, the same kind of solution is available in connection with biology and with sociology, but I will not attempt to give the solution. Particularly in connection with sociology, I think that there are special problems which make the solution more complicated.

First, it is clear that we must dispense with any step-by-step identification of psychological properties first with biological properties, then with chemical ones, etc. The direct interaction of the psychological with the non-biologically physical means that psychological properties must, if we are able to do it at all, be directly identified with properties from physics.

If, however, we have to identify mental properties with complex properties from physics, it is clear, I think, that those properties will have to be disjunctive properties. Now it seems to me that disjunctive properties are perfectly real. The real problem for people who wish to identify mental properties with disjunctive properties consists in making the choice of disjuncts <u>non-arbitrary</u>.

Consider a closely related problem: finding some properties of objects with which to identify redness. For convenience, we will suppose that we are concerned with just one precise shade of redness. It is known that there are a large number of different molecular structures which will reflect light of wavelength around 5000A^{o}. Given this fact, we might feel tempted to identify redness with the power to reflect light of around 5000A^{o}. That has obvious difficulties, however. First, powers, if they are not identical with microstructures, are onto-

logically dubious. Second, the concept of a power, if it involves some kind of de re necessary connection with what it is the power to produce, is problematic. Third, the specification of which power it is runs into the same problem about disjunctions: there are a vast number of combinations of wavelengths that are causally responsible for objects looking red in normal conditions to humans. These difficulties might lead us to identify redness with something in humans, say some property of sense data. But apart from its other well known difficulties, this move flies in the face of the fact that the central use of "x is red" is as a property of physical objects.

We cannot say that redness is identical with many different micro properties of objects, because the transitivity of identity prevents it. We might say that redness is a fiction in that there is no such single property, but it strikes me as counterintuitive to deny that objects are red. So redness must be identical with a disjunctive property.

Now the interesting thing about humans is that they are fairly coarse discriminators of the world. That is to say, humans are so constructed (genetically or by learning) that many quite different states of affairs have the same effect on them. This is why humans cannot discriminate between some objects reflecting quite different light spectra. Machines can be built to distinguish between such spectra even when humans cannot do it.

This coarseness of discrimination on the part of humans is what is responsible for humans having many of the concepts that they do. We give the one predicate "x is red" to quite different objects because we cannot discriminate a difference between them (more precisely,

because they have a similar effect on us and one which is important enough to us, for us to give them names). The appearance of human beings (and biological systems in general) in nature thus introduces a complicating factor in physical systems, in that quite dissimilar physical properties suddenly acquire important similar relational characteristics with respect to humans. It is this similarity in their relations that unifies the micro-properties responsible for a human's seeing red. We might say, for example, that the microproperties responsible for seeing red are similar in that they all cause, in normal conditions, certain beliefs in humans. It seems to me that this is how it is with many of our concepts. Humans approach the world coarsely, and as a result of this many different physical properties are lumped together under the one concept.

This does not seem in any way improper. If we agree that there are disjunctive properties and for that matter conjunctive properties, then it is not difficult to say that there are many more complex properties around than humans have names for. In fact, describing human discriminations as "coarse" is a little misleading. Disjunctive properties exist, but we only respond to some of them. A more complex creature could have the abilities to respond both to very fine differences between properties, and also to very broad categories of objects.

I think that this is what Smart was driving at in "Reports of Immediate Experiences".[8] He describes a hypothetical man who has the concept of snarkhood. Something is a snark if it is either an apple, a helicopter,

---

[8] Smart 1970-1.

the moon, or a book about Plotinus. What makes snarks interesting to us and to the man's psychiatrist is that they have a similar effect on the man: they cause him to stand on his head when he sees them. As Smart points out, snarkhood is an objective property, like many other similar, but idiosyncratic properties. The reason why we should have the concept, is because of the similar interactions between snarks and the man.

The snarkhood example is not as useful as it might be. Snarkhood is only a finitely disjunctive property, or at the very least the disjuncts are completely specifiable. A person might find snarkhood a convincing example of a disjunctive property, but balk at the idea of redness, or a person's total psychological state, including dispositions, as being a disjunctive property because we do not seem to be able to draw a limit around the disjuncts to be included in the property redness.

But why should that be mysterious? The first thing to see is that not any microproperty counts as part of the disjunctive property redness. Some microproperties under normal circumstances cause in people the beliefs that something in front of them is green. Others are causally irrelevant to colour discriminations. So the disjuncts are not wholly arbitrary. The second thing to see is that the inability completely to specify the disjuncts is only, I suggest, an epistemological inability. It derives from the fact that in the present state of science we do not know enough about the causal connections between the microproperties of substances responsible for reflection, the spectra reflected by those substances, and the details of human physiologies and their variations. It is these

facts of which it is necessary to be in possession before we can make any sort of general claim about precisely which microproperties are responsible for the relevant effects on humans. So until we know them, we have no hope of making any sort of exhaustive list of the disjuncts of the property, redness. But, if I am right, there is no reason to despair of there being an objective disjunctive property. That we do not know precisely what the disjunctive property is, does not prevent there being one.

In fact, if we suppose that one day we will be in possession of all these facts about the perception situation, then I think that we can say that in that scientific future there would be nothing in principle to prevent us from giving a quite exhaustive list of the disjunctive microproperty that redness is identical with. No doubt the list would be extremely long; indeed it might be too long for the relatively weak intellectual capacities of humans actually ever to exhibit. But that is not a barrier to there being such a disjunctive property, nor is it a barrier to the possibility that a sufficiently intelligent and scientifically well-informed science might one day exhibit all the disjuncts.

Now I think that precisely the same is the case with mental properties, and also mental states, events, processess, etc. It would be unreasonable to deny that the same mental property can be had by the one person in quite different physiological states, by different persons with quite different neurophysiologies, and in theory anyway, by things with very different chemistries, such as Martians and computers. Different microstates can be responsible for the same causal tendencies in the same conditions. Moreover, human abilities to discriminate the nature of their states are gross. So our mental concepts have falling

under them a broad range of physical properties (always providing physicalism is true). But, as we have argued with redness, the difficulty of specifying the disjuncts of the physical property is an epistemological one only, and so one which does not prevent an identification of the mental property with the complex physical one.

These considerations also provide an explanation of the looseness of "laws" connecting the mental with the mental, and with the physical. Looseness is all we can expect at an historical stage of science where we do not know precisely the physical conditions which are causally operative in a given interaction or class of interactions. On the other hand, it could reasonably be expected that if a precise characterisation of just those properties falling under the concept of pain were forthcoming, laws could be tightened up by making them more complicated.

If these arguments are right, then a classical reduction of the mental to the physical by means of contingent biconditionals expressing property identities is still on. There can be true biconditionals, and mental predicates can be coextensive with physical predicates, contrary to what has been suggested earlier in this chapter.

I conclude, then, that the arguments given for the irreducibility of psychology to physics are unsuccessful. This completes my discussion of the various methodologies open to the physicalist.

Perhaps the major theme to emerge from this discussion is that all the defences to physicalism canvassed leave open the possibility that an investigation of what we <u>know</u> about the nature of our mental states

might be too much for physicalism to accomodate. In the next part of this thesis we will conduct an investigation into just that matter. It will be argued that we can come to know too much.

PART TWO

INTROSPECTION AND PERCEPTION

PART TWO. INTROSPECTION AND PERCEPTION

CHAPTER NINE. INTROSPECTION

1. Introspective Knowledge.

There are many important questions in the general area left undiscussed so far, and many interesting directions in which inquiry could be pursued. In the remainder of this book, I propose to restrict myself as narrowly as possible to just two questions. I will argue that dualism is ~~likely~~ probably true, and I will argue that a version of the representative theory of perception is true.

Indeed, the argument for dualism seems to me quite straightforward. Put very briefly it is this: that when we ask what we are aware (i.e. what we know) in introspection, no materialist account of it is true. Before we can come to this argument, however, we must argue that the attempts at materialism discussed so far must all give an account of introspective awareness.

The first question to ask is: what is introspective awareness? We speak both of "awareness of ..." and "awareness that ...". In my view, the former locution is definable in terms of the latter. The argument, however, will, I hope, not depend on assuming this. I will restrict myself as much as possible to talking only about awareness of the latter sort.

The definition of introspective awareness to be given depends on the prior notion of a <u>mental</u> state, event, process or property. As is usual when one depends on a primitive, it is hoped that crucial examples fall

uncontentiously under the preanalytic idea of the mental e.g. the state of having a red afterimage. Granting this, introspective awareness that ..., occurs when we are aware that ..., where what goes in the place of the dots is a description of a mental event, process, property or state of ourselves. That is, to qualify such awareness as introspective is, as far as we are concerned, only loosely to circumscribe the class of facts of which we have the awareness.

What is "awareness that ..."? At the very least, to be aware that ϕ implies ~~to~~ that ϕ, and that is all I propose to assume about "awareness that". In one perfectly ordinary sense, one is aware that ϕ iff one knows that ϕ. So, to summarise these points, x is introspectively aware that ... iff x knows that ...; and what replaces the dots is a description of some mental process, event or state of x. In a slogan, introspective awareness is knowledge of your own mind. This definition follows the one given by Armstrong in A Materialist Theory of Mind.

This is not the only possible definition of introspective awareness. We might be ~~led~~ led in another direction by the considerations that a person can come to know about some of their mental states, e.g. their jealousy, in a quite ordinary manner, say by being told; and that a person can sometimes know about their bodily states, e.g. changes in the condition of their blood, without these being any evidence that the person goes on, or any accompanying sensation. That is, another sense of introspection might contain the requirement that it be somehow direct and non-evidential.

Our account of introspection does not stipulate any one mechanism for introspection, but there is clearly a large class of cases, indeed the ones most interesting to us here, in which our knowledge of our own minds is somehow direct and noninferential.

The sort of account that Armstrong offers of such knowledge seems to me by and large correct, and I will sketch part of it, for it will be useful later on.

If a person believes that p, then to infer that q from p, and so to come to believe that q, is, for Armstrong, a causal process. It involves as a necessary condition that the belief that p cause the belief that q. That such a causal process be worthy of being called inference requires that other conditions be satisfied as well, but we will not go into this. Similarly, coming to know that q by inference from the knowledge that p, has as a necessary condition that the knowledge that p cause the knowledge that q.

To know that p implies that one truly believes that p. So noninferential knowledge that p, for Armstrong, is true belief that p which is not caused by any other knowledge. It is a reasonable further step to take to say that noninferential knowledge that p is a true belief which is not caused by any other belief. What makes such a belief knowledge, then, if inference, and so justification, is removed from it? At the very least, says Armstrong, the belief must be caused by the state of affairs which the belief is the belief that it obtains. So we arrive at Armstrong's account of the central mechanism of introspective awareness: to know introspectively about one's mental states, in a central class of cases, is to have a (true) belief caused by the mental state without the causal

intervention of any other belief.[1]

2. The Physicalist's Problem.

We are confronted with four attempts at saving physicalism as we have defined it from the threat posed by mental predicates. We might give a successful topic neutral analysis of mental predicates, presumably along Ramsey Sentence lines. We might eliminate mental predicates by showing that they are never instantiated. We might be able to show that mental predicates express properties which are identical with physical properties. We might be able to argue for an adverbialist ontology of mental properties.

All these attempts must, in the end, say what it is that we know when, say, we (allegedly) introspectively know that we have a red afterimage. This is because, firstly, introspective awareness is where all the trouble has arisen in the first place. What we seem to introspectively know, are facts which bear no prima facie resemblance to physical facts, and it is just this that seems to make the alternatives, to analyse or to eliminate, both unpalatable. Secondly, and more importantly, if none of the attempts offers an account of what we ~~are~~ introspectively know, then they leave themselves open to the possibility that an argument will be produced which will show that what we know, and hence what is true, about our mental states, is something that a physicalist cannot accomodate.

---

[1] I omit a complication that the obtaining of the mental state must be "empirically sufficient" for the obtaining of the belief. For a discussion, see Armstrong 1973.

Let us look at this more closely. Should all attempts at a topic neutral analysis break down, one move open to Smart would be, as I have hitherto indicated, to claim that all we know introspectively is that something goes on like what goes on when ... . If this could be established, then though it would not save the "topic - neutral - identity - theory" version of materialism, it would save materialism, at least from the threat posed by what we introspectively know. Smart would be required to deny that there are any mental states of the troublesome sort - after all, they are not "topic neutral" and so if they occur dualism is true. But that, as we have indicated, is not so paradoxical if we can provide some account of what we introspectively know. The revolted intuitions will not be so revolted if they can be made to feel that they were not wholly wrong; that something was right about intuition. We might add, though this might even seem something of an unnecessary bonus, that mental langusge still refers or denotes, as Rorty has suggested.

This sort of move, I suspect, would not be so uncongenial to Armstrong either. He acknowledges the difficulty of providing anything resembling the classical idea of an analysis, but then claims that

> It may still be true, nevertheless, that we can give a satisfactory and complete account of the situations covered by the mental concepts in purely physical and topic neutral terms ... it might still be possible that the account has done justice to the phenomena. [2] (emphasis mine)

This suggests quite strongly that Armstrong would be satisfied to see purely physicalist descriptions of all

---

[2] Armstrong 1968 pp.84-5.

that goes on when we have various mental states, including when we are aware we have them, whether or not we can produce analyses.

Another point is this: that introspection might reveal too much. Indeed, I think that an argument to that conclusion is a necessary part of any dualist strategy. After all, arguments for dualism resting on the unanalysability (in physicalist terms) of mental predicates can always be met by denying that the predicates are instantiated, provided that such a move does not also deny what is known about our mental states. There are various sources of information about the human mind, and one key one is what people know about themselves. This would seem to be the principal flaw in Rorty's original paper. He did not argue that dualism was false, and he gave no account of introspective knowledge. In so doing, he ignored the possibility that what mental predicates "denoted" (in his terms) was not just some state of the cortex.

That there might be such a demonstration is, admittedly, prima facie implausible. Nevertheless, we must grant that it is introspection (especially certain of its forms) that is one of the most fundamental problems for the physicalist. It is where the fuss starts from; that we seem to have knowledge of certain states of ourselves that seem nothing like anything physical. In fact, if this sort of thing did not occur, I suggest that the mind would pose no problem for the physicalist, for mental predicates could then be eliminated.

It is sometimes said that dualist arguments deriving from the content of introspective knowledge are mistaken, for far from its being the case that we know

that our mental states are non-physical, what is really the case is that we merely do not know that our mental states are physical, and if that is all that is wrong, then physicalism is in no danger. Now it seems clear that, at least for awareness of a fairly direct (noninferential) sort, we are introspectively aware neither that our mental states are physical nor that they are nonphysical. But of course a dualist argument might not rely on any mistake as crude as that. It might take quite complicated inferences from apparently innocuous facts that we do know, to establish dualism. Or it might be that what we know is not something which implies dualism by itself, but needs some known facts about the brain as well. (c.f. Ch. Three, Eleven.) Either way, the dualist argument cannot be so lightly brushed aside.

Even if the topic neutral analysis were to succeed, physicalism has not necessarily been saved. If we can always retreat from unacceptable mental predicates by denying their instantiation, then we can always retreat from acceptable mental predicates in the same way. If mental predicates are dualist-loaded and because of that people have been inaccurately reporting and describing their states all this time, we must at the least accept the possibility of topic neutral reports and descriptions of our states being inaccurate also.

Someone might reply to this point, by asking how we should determine that topic neutral analyses of our mental states are inaccurate, if all the language we have to describe them is ex hypothesi topic neutral? Surely, it might be said, once topic neutrality could be established, Ockham's Razor and other Principles of Method are enough? We suggested as much in Chapter Three, when

we pointed out that as traditionally conceived, the Identity Theory rested on just two platforms, topic neutrality and Ockham's Razor.

If I might be permitted some philosophical polemic, too much English-speaking-philosophy has become infected with Carnap-ism. If we think that for every genuine philosophical problem, there is an equivalent formulation in terms of language, in the meta-language, then we might find it easy to believe that the philosophical task is complete once the linguistic thesis (in this case, the thesis of the topic neutrality of mental predicates) has been established. But it is precisely one of Rorty's contributions, to note that the failure of the linguistic thesis does not have to be accompanied by the failure of materialism; that there might be considerations which force us to materialism and allow us to hold it in spite of that failure. In short, we must look at the facts as well as our language for the facts. In passing, this is in no way to deny the methodology of the Final Theory in Chapter One. There, we defined physicalism as the doctrine that the final theory contained only predicates from P. Now if Rorty is right and humans really are physical, then (as far as mental states are concerned) the final theory need only contain predicates of P. Physicalism is not the doctrine that the final theory must be expressed in the physicalist predicates <u>of our present language</u>. Conversely, a dualist who holds that the final theory must contain predicates not in P, does not have to maintain that our present language also contains those predicates. Our present language, in containing only topic neutral predicates for mental states, might be impoverished.

There is another, and somewhat incompatible (in method but not in final result) function that an investigation of what we introspectively know can serve. This is, that if we find that what we introspectively know is not compatible with physicalism and also that we do have predicates for expressing it, then we will have established the non-topic neutrality of certain of our mental predicates. Introspective investigation, then, can be a part of the investigation of topic neutrality. The point is even stronger than ~~that~~ this, for if it is a matter of putting together some of the things that we know, or can know about our minds, then we not only establish ~~the non-topic neutrality of mental predicates~~ that mental predicates are not topic neutral, but also ~~the applicability of the predicates~~ that the predicates are applicable, contrary to Rorty.

Substantially the same points can be made against someone who opts for the contingent identification of mental properties with physical ones. We should ask such a person why they wish to identify the properties expressed by mental predicates with physical properties. We should ask such a person why they think that such properties <u>can</u> be identified with physical ones. As has been argued, the wish to identify mental properties with physical ones derives in part from Ockham's Razor. After all, as we have already noted, physiology is not particularly complete. Thus, for example, someone who argues

1. The property, having a red afterimage, is the effect, at t of stimulus S.
2. The physical property, p, is the only property which is the effect, at t, of stimulus S.

∴ 3. Having a red afterimage = p

clearly commits themselves, in (2) to some version of Ockham's Razor or related Principles of Method. But

other things might not be equal, and there might be a reason to believe in the more complex theory. Again we note that if there is going to be an argument for dualism, introspection would seem to be the principal candidate for a place for it to come from. In any case, if such an argument is forthcoming, then it is an argument presumably for, among other things, the denial of (2) above.

3. What is Introspected?

How shall we find out what we know about our mental states? One place to start is to ask whether we know anything at all about them. The answer I want to give is that it does not matter for our purposes if we do not. More accurately, it does not matter if we do not know anything about our afterimages, pains, and the like. In particular, it does not matter if we do not know whether we have them. This point is principally aimed at sidestepping Rorty. I do, however, want to focus on those occasions when we allegedly have them (those occasions about which there is a dispute as to their correct description). Perhaps we might define those occasions as occasions on which we believe we have particular mental states (and, often enough, report them).

I want to ask of these situations: what do we know of ourselves when they occur?

Perhaps in introspection we "imperfectly apprehend" our brain states. As Keith Campbell puts this suggestion[3], it amounts to distinguishing between how our brain states "appear" or "seem" to us, and how they really are.

---

[3] Campbell 1970 pp.105-6.

We are not aware of our pains "as" brain states, but "as a condition which hurts" (p.105). We "grasp" them "in the guise of the painfulness of the pain" (p.105). The key point for Campbell is this: that appearance is not necessarily reality. If something only seems a certain way, then there is no need to suppose that it really is that way; appearance is "ontically neutral"(p.106).

What, according to this view, do we know or believe when we (allegedly) have an afterimage or a pain?[4] Words like "appear", "seem", used when talking about our knowledge when we have mental states, are not clear. The most plausible interpretation seems to be this: our brain states seem to be pains in that we believe that we have pains. Now how is it that our brainstates could seem to us as pains? It cannot be that we believe that our brainstates are pains, or painful, for it is surely false that we do. A helpful analogy might be with a person seeing a tree in dim light and taking it for a fox. We might say that the person took the tree for a fox, that the tree appeared to be a fox, that it seemed to be a fox. We would not say that they believed that the tree was a fox. Why we should say these things, it seems to me, is because of a story about the causal role that the tree played in the mistaken belief. Similarly (presumably)for Campbell , we ought to say that we take our brain states to be pains, to be painful, in that we believe that we have pains, and it is brain-states of a certain sort which play a special causal role in the production of that belief.

---

[4] The "allegedly" will be dropped hereafter except where it is necessary to avoid confusion.

It seems to me that this position can be shown to have to deal with the considerations already raised about analysis, property identification, etc. This is so for the following reason. Either the belief that we have a pain is true, or it is false. If it is true, then the imperfect apprehension theory must be able to say what it is to have a pain and what it is to believe that we have a pain. In particular, it must be able to guarantee that having a pain is not being in some physicalistically unacceptable state. If on the other hand it is false, then we will need a reason for thinking it false, and some kind of account of which true introspective beliefs we have in order to cushion the blow to our intuitions: in other words, we will have to defend eliminative materialism. But with either of these consequences, it has already been argued that we need to look at the question of introspection. In short, the imperfect apprehension theory insofar as it tells us that we believe that we have pains and that this belief is caused by brainstates, is the beginning of an investigation about what is introspected, but much more needs to be determined before we can rest easy with physicalism.

The same argument can be given to a person who thinks that all the dualist fuss arises from the simple fact (consistent with physicalism) that we are merely not aware of our mental states as physical. For what could "aware of ... as ... " mean here? To be aware of a as being F is only, as far as I can see, to know introspectively that a is F. But to tell us this small amount about our introspective beliefs is not a great deal of help when investigating what their content is.

Campbell's own criticism of the imperfect

apprehension theory is worth noting. He says that "seemings" might banish painfulness to the realm of the merely apparent, but then we should be left with a class of irreducible seemings.

This is true enough, provided that you do not propose to analyse "seems", "appears" in these contexts in the way I have suggested i.e. using beliefs. You would worry about whether "x seems painful to y" is a member of P. You certainly would not have avoided the sort of problem that "x has a pain" raises. You would just have pushed it back.

Well, then, what do we know when we are alleged to have a red afterimage? It seems clear that on some of these occasions we can tell that we are in a similar state to the state we are in on other such occasions. It is tempting to say that knowing that we have a red afterimage involves, at least on some such occasions, knowing that we are in a state similar to certain other states. Smart certainly has said as much. But to say this would be to make too much of an assumption against Rorty: namely, that on such occasions we do know that we have a red afterimage, and hence that "x has a red afterimage" is instantiated.

Even if we say something as weak as this: that on such occasions we believe that we have a red afterimage, then we must agree that we are in a similar state (i.e. the belief state) on such occasions, and it is clear, I think, that <u>on those occasions</u> we know this much. If, then, we agree that for some of our mental states (i.e. beliefs) we sometimes do know that they are similar to one another, there would seem to be no reason to deny the undoubtedly strong intuition that often

when believing we have a red afterimage, we do know that our state is similar to certain other states; even if we do not say that the state is the state of having a red afterimage.

It must be made clear that "similarity" will have to be interpreted fairly loosely here. No doubt no two of those occasions when we (allegedly) have a red afterimage, are occasions on which we have exactly similar brain states, (even exactly similar states in the visual centres). But it would be unreasonable to use this fact to cast doubt on the fact that we know that we are in similar states on such occasions. If physicalism is true (and even if it is not!), our introspective discriminations of our states are evidently gross, not fine, and if introspection is discrimination of physical states then it will be discrimination of relatively global features of those states. Thus "similarity" will be "similarity in some respect" (which is not to imply either that there are respects, or that we can tell in which respects our states are similar).

With these hurdles aside, let us agree that often enough when we (allegedly) have red afterimages, we know that we are in some state similar to other states which we have. Which other states? Obviously not only those states had on other occasions when allegedly we have red afterimages, unless we know something more about the states. (Otherwise there would be nothing to distinguish what we know when we have a red afterimage from what we know when we have a green afterimage.) In fact, of course, the state we are in when we have a red afterimage is similar to the state we are in when under normal conditions we really see something red; and we know this

much, often enough, at the time.

If that is what very often we know on such occasions, nevertheless there is some reason for thinking that it is not what we know on all such occasions. For there are imaginable cases (which no doubt even occur) where the suitable conditions of alleged occurrence of red afterimage obtain, we believe that we have a red afterimage, and yet it would seem to be the case that we neither know nor believe that we are in a state like the state when under normal circumstances we see a red object. It is not difficult to imagine that you have several red afterimages before seeing anything red, or even never see anything red at all. In such circumstances, it can be true that you believe that you have a red afterimage, but not believe that the state you are in is similar to the state you are in when under normal conditions you see something red.

Would we really want to say of such a case that you believe that you have a red afterimage? After all, how could you not believe that it is red, without also believing that it is like what you get when you see a red object?

This is a somewhat tricky area. It might be that this is a place where the analysis of introspective awareness in terms of knowledge and, ultimately, beliefs, breaks down. Surely we would want to say that in such a case we could be fully aware of having the red after-image; fully aware of all its aspects, particularly its redness. If it turns out that we cannot say that we believe that it is red, then perhaps we have to complicate the analysis of awareness. But after all, would it be so unusual to say of the case in hand that we believe that

we have a red afterimage? One could imagine, pace Wittgenstein, a person coming to acquire the concept of redness, and the word "red", just from their afterimages, hallucinations, etc. (and acquiring the counter-factual ability correctly to describe pillar-boxes), or even coming to acquire it by having a friendly scientist stimulate their brain. So a person might definitely believe that they have a red afterimage, but have no opinions on whether they are in a state similar to the state they are in when under normal conditions they see something red.

A reply might be this: that such a person could not believe that the afterimage is red without having the concept of red, and they could not fully have the concept of red without believing, at the very least, that they are in the state that they would be in if they were to see something red.

Such a reply would meet the objection sometimes given to the Smart-like account of what we are introspectively aware, that it cannot account for the possibility that we have just one "experience of red". The reply contends that when we believe that we have a red afterimage, we believe that we are in that state that we are in if under normal conditions we see something red, and construe the "if"-clause as neutral as to whether its antecedent is ever satisfied or not. But we could not rest easy with this reply, I think. First, and perhaps this is not a very strong point, we could hardly claim that our knowledge in those circumstances derives from, or is based on, any knowledge of the actual similarities between our states. One might feel inclined to ask how it could be that one would know which state it would be that is like what would go on if we were to see something

red. After all, this account does not tie down the intrinsic nature of the state at all: it leaves open the possibility that we are in various sorts of states when we have a red afterimage and when we see something red. For instance, it would seem to be consistent with the possibility that one is having a grey afterimage, and, because one is believing that one has a red afterimage, is supposing that this is the sort of state that one would also have if one were to see something red.

I am unsure as to the strength of the preceding argument. The second argument, however, is certainly strong enough for my purposes: if in believing that we have a red afterimage, we believe that our state is like what it is or would be if we see something red, then we need to know what our state is, and what we believe about it, when we see something red. There is no doubt that many people often do see red things, and saying that having an afterimage is like seeing something red in no way guarantees that what we know when we have afterimages is physicalistically acceptable, until we decide that seeing red things is acceptable too. What follows the "if" in the account must be supposed to obtain on occasion. So to satisfy physicalism, we shall have to be satisfied that seeing something red is harmless.

4. Introspecting the State of Seeing.

What, then, do we know about the state that we are in when we see red objects? Clearly, there are many situations when we are looking at red objects under normal conditions, seeing a red object and so knowing that there is a red object in front of us, when we also are not aware that we are in any particular state. We

are absorbed in our surroundings, as it were. There is a difference between seeing something and being aware that we are seeing it.[5]

Knowing that we are seeing it is knowing that we are in some state, the seeing state. It is clear, I think, that when we know that we see a red object, often enough we know that we are in a state similar to other states that we have on other occasions.

One similarity between the seeing-a-red-object states, is that they are caused by, or occur simultaneously with, there being a red object in front of our eyes in normal conditions. Another similarity, is that when we know that we are in such states, we typically know that they are states caused by, or occurring simultaneously with, there being a red object in front of our eyes in normal conditions. That is to say, we typically know <u>that</u> we are in the state caused by or occurring simultaneously with, there being a red object in front of the eyes.

We know that such states are similar. We know that we are in a state, such states being caused by red objects. Is it that the similarity which we know to be between such states, is just the similarity of being caused by red objects? That is to say, is the similarity that we know to be between our seeing-a-red-object states,

---

[5] If questioned, we almost invariably reply that we see it, thereby presumably confirming that we invariably know we see it as well as just seeing it. But this is easily explained (1) by pointing to the fact that the question "Do you see the red thing?" (as opposed to "Is there a red thing there?") often has the effect of drawing our attention to the seeing of the thing, and (2) by the fact that the question "Do you see it?" and the answer "I see it" are often treated as "Is it there?" and "It is there".

only similarity in the respect of being caused by a red object? Or do we know them to be similar in other respects as well?

There does not seem to me to be any a priori difficulty in our having such knowledge about our states and no more, always provided we can get an Armstrong-style causal theory of knowledge going. What could the knowledge that we are in a state, the nature of which we know not, which has the relational property of being R to a, amount to? Surely nothing but the (true) belief that we are in some state such that it is $\hat{R}$ to $\hat{a}$, with the appropriate Armstrong-style causal conditions attached to our acquisition of that belief. There can surely be no objection to our believing some proposition of the form: (Ex)xRa. We believe such propositions all the time.

Nevertheless, this is not all we can know about the similarities between our states. It is clear that one of the respects in which our seeing-a-red-object states are similar to one another, is that respect in which we have already noted that the states occurring when we allegedly have a red afterimage, can also be known to be similar. Our allegedly-having-a-red-afterimage states, are similar to one another and to the state we are in when under normal conditions we see something red. We have already noted this, just as we have already noted that we often enough know these facts at the time, as it were (directly, we might say). Similarly, we know that our seeing-a-red-object states are similar, and furthermore that one of the respects in which they are similar, is also a respect in which allegedly-having-a-red-afterimage states are similar to them.

From which it follows that it is not the case that the only respects that we know that our seeing-states are similar to one another, is that they are caused in a certain way, or occur in certain conditions (when there is a red object in front of the eyes, etc.). For these conditions are absent when we have a red afterimage !

This argument shows that we know more about the similarities between our seeing-states, than merely their similarities in respect of which causes they have. The next point I want to make is that neither the knowledge of their similarities, nor the knowledge of their causal antecedents, is enough to account for our knowledge of the differences between our seeing-states.

5. Knowledge of Differences Between Our States.

First let us consider just the proposition that when we have a red afterimage, what we know is just that we are in some state caused (or occurring when etc.) in a certain way. Now I claim that we can also recognise that our states had at such times are different from one another. The state that we are in when we (allegedly) have a red afterimage is different in a crucial respect from the state we are in when we have a green afterimage. Furthermore, I claim that we very often know this fact - can tell it at the time, as it were. The same pair of points goes for the states we are in when we see a red object and a green object respectively.

Now the difference that we know to be between our seeing-a-red-object states and seeing-a-green-object states, is not just a difference in respect of having different causes. The reason for this is as follows.

Nothing in logic guarantees that a green object in certain favourable conditions will not cause precisely the same state as a red object in certain favourable conditions. The normal human physiology (or spiritual physiology) might just be constructed so that no difference is discriminated between red and green objects - that red and green objects make no difference to the states of the discriminator. Indeed, this would appear to actually be the case in at least some cases of red-green colour blindness.

Suppose someone should reply that the states would still be different in respect of one sort being caused by red objects, and the other sort being caused by green objects. Is not this a difference, it might be asked, and could it not be the case that in discriminating our seeing-a-red-object states from our seeing-a-green-object states, what we are discriminating - what we know - is that one sort of state has one sort of cause and the other sort of state is different in that it has a different sort of cause?

But this is to miss the thrust of our argument. The reply depends on the assumption that being caused by a red object is a genuine difference from being caused by a green object. But this could ~~only~~ necessarily be the case, only if "x is caused by a red object" necessarily has a different extension from "x is caused by a green object" and there is no guarantee that this should be the case. Our argument is not just an argument that red objects might cause qualitatively the same perceptual states in human beings, but more strongly that they might cause exactly the same states - a plurality of causes for each such state.

Now this of course does not occur. But the point is that unless we claim that it necessarily does not occur, then we cannot rely on apparent differences of causal relation to guarantee us differences in the states we discriminate in ourselves. So we must conclude that at least some of the differences that we know to be between our states, are not differences in respect of causal properties.

Another argument establishes the same conclusion. We can tell differences between our alleged-afterimage-states as well as our seeing-states. We can know that the states when we allegedly have a red afterimage are different from the states when we allegedly have a green afterimage. But these differences are not differences in respect of the first sort of state being caused by red objects and the second sort of state being caused by green objects, for neither sort of state has such causes. In fact, it is clear that one respect in which we can know that they differ is the same respect in which we can know there to be a difference between seeing-a-red-object states and seeing-a-green-object states. But the first difference is not a difference in respect of allegedly-having-a-red-afterimage states being caused by red objects and allegedly-having-a-green-afterimage states being caused by green objects. Neither, therefore, can the second difference be. That is to say, we know that our seeing-states are different, and we know more about their differences than merely that they are different in respect of having different causes of a certain sort.

If we are clear on this point, then we should be clear that when we add the <u>additional</u> knowledge that our states are similar in various ways, then the combination

of knowledge about similarities and knowledge about causes is not enough to account for the knowledge about differences. We know at least three sorts of independent propositions about our states. We know propositions of the form "x is caused by y", of the form "x is similar in some respect to y", and of the form "x is different in some respect to y". Let us spell out the argument for this claim in some detail.

6. The Independence of Knowledge of Differences, Similarities and Causal Relations.

In a slogan, the argument is that similarities are insufficient to determine differences. We know (1) that our states had while allegedly having a red afterimage are similar to one another, and to the states we are in when seeing something red in normal conditions. (2) Similarly for the having-a-green-afterimage states, and the seeing-a-green-object states (3) that our seeing-a-red-object states are similar in respect of having similar causes. (4) Similarly for our seeing-a-green-object states. Now are these four sorts of knowledge enough to guarantee the fact that we know (5) that our seeing-a-red-object states are different from our seeing-a-green-object states, and (6) that our having-a-red-afterimage states are different from our having-a-green afterimage states? The answer is no; for the same reasons as given above for the conclusion that (3) and (4) are insufficient to determine (5) or (6). First, because knowing (1), (2), (3) and (4) does not entail knowing either (5) or (6). Second, because one of the respects in which our seeing states differ and are known to differ, is also a respect in which our afterimage states differ and are known to differ, and this is not the respect of one being

caused by red objects, the other being caused by green objects, for the afterimage states in question are not caused in this way.

7. Knowledge of Effects.

We now add the complicating factor of the effects of our states. We do very often know that our seeing-states and afterimage-states are causally operative in a given situation. We ask the two questions (a) Can this sort of knowledge be accounted for by any combination of (1) - (6) above, or is it independent? (b) And if it is not entailed by any combination of (1) - (6) above, can it in combination with some of (1) - (6), be used to account for other of (1) - (6) ?

The first question barely needs stating in order to be answered. Obviously, knowing about the causes, similarities and differences of our states gives us no information about their effects.

As to the second question, I propose to restrict myself to the particular question of whether our knowledge of the differences in the actual effects of our states is sufficient to account for all the known differences between our states. It has certainly been Armstrong's view that since our mental concepts are primarily concepts of causes of various sorts, differences in the effects of our mental states are what primarily distinguishes our mental states (regarded as kinds) one from another. If we believed this, we might think it plausible that our knowledge of the differences between our states amounts to knowledge of the differences in their effects. Furthermore, there is a crucial difference between (1) accounting for the known differences between our states in terms of

(known) differences between their effects, and (2) accounting for the former known differences in terms of differences between the causes of the states. This is, that while difference between the causes of the states does not guarantee difference between the states, difference between the effects of the states, under standard conditions, does guarantee difference in the states. Causality is such that the like causes in like conditions give like effects; so different effects, given that conditions remain the same, guarantee different causes.

Now mere differences in the actual effects of our states does not guarantee differences between our states. This is because our states do not by themselves causally determine their effects, but only in conjunction with surrounding conditions. But this is not enough to establish what I want to establish. In order to show that some of the knowable differences between our states are not differences in respect of their effects, there needs to be a case in which we can know that two states, say an allegedly-having-a-red-afterimage state and an allegedly-having-a-green-afterimage state are different, and yet they not have different effects. But this can certainly be the case, even if we suppose that those states, in belonging to the extensions of different predicates, have different characteristic effects. In different surrounding conditions, having a green afterimage and having a red afterimage can and sometimes do have identical behavioural consequences. Indeed, this can be the case independent of what the causes of those states are.

It might be replied that having a green afterimage and having a red afterimage can plausibly be supposed

always to have different effects, even if those effects are not behavioural effects, but, say, neurological effects. But it is difficult to see what advantage for physicalism there would be in saying this. If we do not know what the neurological effects of our states are, only that those effects are different, then why not simply say that we can know our states themselves to be different in ways other than differences in respect of effects, and claim that the differences that we know to be between our states are neurological differences, although we do not know this?

Furthermore, the suggestion of the previous paragraph would also run into the difficulty that not only would the neurological nature of the neurological effects of our states typically be unknown, but also their status as effects would be unknown. It is surely false that we typically know in any direct way anything at all about the neurological effects of our states, even that there are neurological effects at all.

To cut a long story short: I claim that at least some of the knowledge of the similarities and differences between our states is knowledge of similarities and differences other than the (known) similarities and differences between their causes and effects.

There is a particular case of this problem that arises when we move from considering the <u>actual</u> effects of our states to the <u>tendencies</u> that they have to produce their effects. It might be thought that since different mental states typically have different causal tendencies (for if they did not there would be a problem of why we should have different terms for them), knowledge of difference in causal tendency is enough to account for our

knowledge of the difference between, say, having a red afterimage and having a green afterimage. In the next section, we will look at this problem.

8. Knowledge of Causal Tendencies.

For Armstrong, the mind is a field of causes, but we do not have to suppose that, on every occasion, a mental state produces its characteristic effect. Sometimes mental states only tend to cause their characteristic effects, or are apt to cause them. Furthermore, for Armstrong, what individuates mental states is their aptness to cause differing, loosely specifiable, behaviour. Now there is no doubt that we do sometimes know that we are in a state which is tending, unsuccessfully, to cause behaviour. For instance, we can resist the effects of (or allegedly when, pace Rorty) having a pain. Can we say then that knowing different tendencies of our states is enough to enable us to account for all the differences we know to be between our states (leaving aside differences in causes, which we can suppose that we know nothing about in particular cases)?

In sections 9-11 of this chapter, it will be argued that there are certain properties of our afterimage-states and seeing-states which we can know them to have and which form the basis for ascribing one set of differences to those states. In section 12, and in the next chapter, it will be argued that those properties are not to be identified with causal tendencies. It will be argued particularly that our states can be known to have those properties when the most plausible candidates for causal tendencies for them to be identical with, are absent. If that argument is correct, then the problem raised in

this section is solved in the negative. Some to the knowledge of the differences between our states is not mere knowledge of difference in causal tendencies.

One point should be made here, though. It might be said that if two states can be shown to be different, then they thereby have different causal tendencies, namely the respective tendencies to produce beliefs about their nature in us.

There are two answers to this. The first is that from the fact that two states can be known to be different, it by no means follows that they invariably tend to produce that knowledge. Introspection is a learnable art, (this point will be amplified in the next chapter) and the learning to practice it involves coming to have the ability to gain beliefs about the nature of our states. There is no necessary tendency present always to have those beliefs. The second answer is that if the tendencies to produce beliefs about the nature of our states are to be used to determine a difference in the tendencies of those states, then this can only be done if the knowledge (beliefs) produced are themselves different from one another. But this could only be so if the beliefs apt to be produced have a different content. But this implies that what it is that is believed about our two states is different, and therefore, if those beliefs constitute knowledge, then we have knowledge of differences in our states other than differences in respect of their tendencies to produce different beliefs. The alternative is a regress: that our states are different in that they tend to produce beliefs which differ in the respect that they are beliefs that our states differ in that they tend to produce beliefs which differ in that ... .

In the next section, I will proceed with the argument just now outlined.

9. Knowledge of Respects?

Smart has claimed in a number of places[6] that we can tell that the state we are in when we have a red afterimage is similar to the state we are in when we see something red, without being able to tell in what respect they are similar. The preceding argument, however, establishes that we have quite a bit of information about the respects in which we know our states to be similar. We now know that the states are (sometimes) known to be similar in respects other than the causal respects in which they can be known to be similar. And we know that they are (sometimes) known to be different in respects other than the causal respects in which they can be known to be different. This conclusion is reinforced when we note that the state we are in when we are allegedly having a red square afterimage is similar in some respect to the state we are in when we allegedly have a green square afterimage, and similar in a different respect to the state we are in when we allegedly have a red round afterimage, and that we can know these facts at the time as it were. (If they were similar in the same respect, then it would have to be the same respect in which the having-a-green-square-afterimage state was similar to the having-a-red-round-afterimage state, and there is nothing necessarily similar between these two states - save that they are both having-an-afterimage states, but this is not the original respect we had in mind.) A similar point can be made about

---

[6] e.g. Smart 1959, 1963a.

the differences that we can know there to be between the three states.

The account of similarities and differences that I favour, is that respects are properties. Similarity in some respect, is sharing a common property. Difference in some respect, is one thing having a property that the other lacks. Knowledge of similarities and differences without knowledge of the respects, is knowledge that there is a shared property, or knowledge that there is a property had by the one and lacked by the other, without any knowledge of which properties they are.

But how can there be knowledge that there is a shared property, and that this property is not the same property as another shared property, without the knowledge of which properties they are? That is, without the knowledge that the shared property is the property p, where "p" uniquely names the property in question. After all, we are supposing that we can tell sameness and difference of respects. And this amounts, in property language, to being able to tell of any property (respect) whether a given property is identical with it or different from it. What more could we need to be able to know which intrinsic properties our states have?

Another consideration inclines me in the same direction. In introducing the possibility that we be able to tell the effects of our states as well as the causes, we implied that we be able to tell that a given state be the <u>same</u> state which is both the effect of C, and the cause of E. But how shall we be able to tell that, if all we can tell is that we are in <u>some</u> state which is the effect of C, and <u>some</u> state which is the cause of E? Similarities in <u>some</u> respect will not guarantee it (though

differences in some respect will defeat it, and hence absence of difference in some respect will guarantee it). But how shall we know that without at least knowing that the effect of C is the same in all respects as the cause of E i.e. without having some way of knowing, for any property of the effect of C, whether it is the same or different to any given property of the cause of E?

These questions might seem excessively rhetorical, so I propose to argue that we have more information than even this complex system so far sketched of known similarities and differences between respects can account for.

10. Knowledge of Certain Other Properties of Our States.

Suppose that we are in a given kind of state just once. Suppose that unbeknownst to us, it is like the state that we would be in if we were to see something red. (It is unbeknownst, because we can suppose that we have no knowledge of the causes of the state; suppose for example that it is produced by a probe but we do not know this. We can of course suppose that we know that it is a "visual" state - the sort of state caused by looking at things in normal conditions rather than by listening to them.) The case may seem far out, but perhaps only because of the choice of red. There are no doubt many instances of colours seen just once or not at all.

We can also suppose that in the case we have no knowledge of the effects of the state (leaving aside for a time those effects which may be tendencies). So we have no knowledge of causes, effects, similarity of

causes to causes of our other states, similarity of effects, difference in causes or effects. Nor is it the case that we know that the state is similar to other states (in the respect in which having-a-red-afterimage states and seeing-red states are all similar to one another). For there are no such similar states that we ever have.

Now we are aiming to show that there is more to be known in the given example than the so-far delineated similarities and differences. But someone might think that in having a single instance of such a state, we do not guarantee any knowledge of it (being in a state and knowing we are in it are distinct, as we have already said). Thus how can a case like the above begin to show that there is anything further to be known about our visual states than the normal similarities etc.?

All I want to say about this objection is that it seems clear that we can imagine that in the case in question we know that the state we have is different from any other state we have. In real life examples, we can surely often tell that we have never seen that colour before. There is no difference in principle if the state is produced by a probe.

Now somebody else might think that all there is known in the case is that the state is different from other states we have had. They might concede the point of the previous paragraph and concede no more.

The reply to this is that there is something more to be known about our state. Perhaps this can best be brought out by complicating the example.

Case 1: Suppose that as well as there being a kind of state that we have just once, there is another kind

of state (say, a state like that which would be caused by seeing something green under normal conditions) which we never have. In addition to this case, consider a second case as well, case 2. In case 2, everything is exactly as before, except that it is the second kind of state which is had just once, with the knowledge as before that it is different from any other of our states, with appropriate lack of knowledge of its causes and effects, and it is the first kind of state is had not at all. Now I claim that there can be (if we are paying attention!) a difference in what we ~~are~~ introspectively know in the two cases. Indeed if there were not a difference in what we are aware, I cannot see how there would necessarily be any difference between the cases at all, (for difference in causes, even hypothetical ones, does not guarantee difference in the states, as we have already pointed out). And it seems clear to me that the cases are different in a respect that we can be conscious of.

Do not confuse what I am saying with the claim that we know that our state is different in some way from other states. I am claiming that the content of our knowledge of our state, i.e. what we know about our state, is different in the one case from what it is in the other. But if this is the case, then it cannot be true that all we know about our state in the (first) case is that we are in a state different from any other state that we are in. For we also know that much in the second case, yet we have conceded that there are differences in the content of what we know between the two cases.[7]

---

[7] Perhaps someone might think that in case 1 we also know that we are in a state different from the state we would be in were we in the hypothetical case 2, and that this determines a difference because the same

Somebody else might think that there is something extra to be known in the example, but that it can only be known by someone who "has the concept of red". Until you suppose that in case 1 we have the concept of red, or know the meaning of "red", it might be said, then you cannot describe a difference in what is known between the two cases. But if you cannot describe a difference in what is known, then you have no reason to believe that there is a difference in what is known. But once you do allow that we can use "red" in case 1, then you can say that what extra we know is that we are in a state like the state we would be in if we were to see something red. And this will be enough to determine a difference in what is known between the two cases, for in the second case, we know that we are in a state like the state we would be in if we were to see something green.[8] The point of the objection is this: that this further counterfactual knowledge is acceptable to the physicalist.

---

cannot be said of case 2. But it can be supposed that in case 1 we have no idea of what the unknown state of case 2 is like, and that in particular, we have no guarantee even that with a different imagined stimulus (which we know nothing about in case 1) we should get a different state.

[8] It should not be thought that this objection must fail because in order to have the concept red, we need previously to have seen red things (or been in states appropriately like those had when a red thing is seen), and hence that we cannot suppose both that we are in that kind of state just once and that we have the concept red. There is no learning history which is logically necessary for someone to have any concept at all, for there is no contradiction in supposing that Aphrodite emerge full-grown from the waves with all our language, concepts and abilities to describe.

However, even granting the doubtful point that we do not have words to describe the difference, there is no reason to think that we cannot describe a difference between the knowledge of our states in the two different cases, if we do not have the concepts red and green. We can, after all, say that in the one case, we know that we are in a state with this property (or in a state with this respect, or in this sort of state), and in the other case that we know that we are in a state with that property. What defines a difference in what is known, is that this property is not the same property as that property. (We do not have to know in the one case that there is a difference in what we know between what is the case and some hypothetical example; we do not have to know that this property is not the same as some other property of which we have no knowledge.) And indeed this does seem to me precisely the difference in what is known between the two cases. If somebody has never heard of chartreuse, or "chartreuse", or seen anything chartreuse, and you stimulate their brain to the state that chartreuse gives, they would certainly not know that it is like what chartreuse objects cause, but they would still in theory be in a recognisable state, and still able to say "It is like this". This is just the sort of knowledge that new experiences carry with them. In no sense is it necessary that concepts precede experiences; in fact it is often because of new experiences that new concepts emerge.

Now I want to say that it is these properties of our states which distinguish the state we are in when we allegedly have a red afterimage, from the state we are in when we allegedly have a blue afterimage, and which form the basis of our knowledge of the differences and

and similarities between our allegedly-having-an-after-image states and seeing states. Far from its being the case that we merely know that our states are different and similar in some respects without knowing about the respects, rather is it the case that we have quite extensive knowledge about the respects. We tell similarities because we know that they both have this respect, and tell the differences because we know that this state has this property and that that state lacks it.

Why should anyone believe this claim? I think that there is only one way of getting people to accept the claim, and that is to draw their attention to the facts. When we know that it is this property rather than that, the property that we know it to have is precisely the property that we know to be in common between having-a-red-afterimage states and seeing-a-red-object states, and know to be lacking in such states as having-a-green-afterimage states and seeing-a-blue-object states. In the former mentioned states, we often know that our states have this property, and it is on the basis that they share the property, that we ascribe similarities to them.[9]

11. Sensations-Of-Red.

Perhaps the reader is impatient with "this property"

---

[9] This is not to say that we do not on occasion, or even quite often, ascribe similarities (which are in fact similarities in the respect in question) without knowing the respect (property) in common. Nothing said so far implies that. Nor is it to say that we necessarily infer that our states are similar from the fact that they all have a given property. If we say that it is on the basis that they both have p that we know them to be similar, this is not the same as saying that it is on the basis of our knowledge that they both have p that we know them to be similar.

and "that property", and would like predicates or names for them. Indeed it might seem strange, if we do know frequently that our states possess a given property, that the language we speak does not contain terms for such properties. I should like to defer answering this worry fully until later, with the comment that it is not so obvious that the language does not. However, it will be convenient to have some terms, rather than talking about "that respect common to allegedly-having-a-red-afterimage states and seeing-a-red-object states, and lacking in allegedly-having-a-green-afterimage states and seeing-a-blue-object states". So we will have the predicate "x is a sensation-of-red", true of our states just when they have the above property. Similarly we will say that in such cases our states are sensations-of-red, and that we have sensations-of-red. The relevant property of our states, will be that of being a sensation-of-red. The hyphenation is there to warn that there is no suggestion that states having the property in question are sensations, or are red. We choose "red" in the predicate, however, for a good reason: that it is obvious that sensations-of-red have something to do with seeing red objects, and with having red afterimages, and both these descriptions have "red" occurring in them. There are corresponding predicates "x is a sensation-of-green", "x is a sensation-of-blue", etc.

Now that we have these terms, we can say that we, the readers of this thesis who have swallowed the argument so far, now know that we have sensations-of-red ("have" sounds better than "are in"). Sensations-of-red are generally complex things, being states which have similarity properties and difference properties in other respects that we can know about, and, though we

will not pursue the question here, other properties which we can know the states to have.

While we have determined that we know that our states have properties which we can name with "being a sensation-of-red", "being a sensation-of-green", etc., we have not yet determined whether these properties are acceptable to the physicalist or not. For all we know at this point, the physicalist might easily accomodate them. So we will turn to this problem next. Before we do, it is worth mentioning in passing that a Smart-Lewis-style topic neutral analysis would seem to have been defeated by what we have said. The reason for thinking so is as follows. Accept for the purposes of discussing Smart that we do have afterimages, identify the state of having an afterimage with the state had when we (allegedly) have afterimages, and ask what the analysis of "x has a red afterimage" is. I claim that essential to any such analysis is something which caters for the property, being a sensation-of-red. This is because if we remove that property from the state of having a red afterimage, it ceases to be the state of having a red afterimage and becomes the state of having a green afterimage, or the state of having a blue afterimage, or ..., (or not the state of having an afterimage at all, if we do not replace it with anything responsible for "after-image colour"). But, then, no aspect of the analysis catering for similarities to seeing-red states, differences from other such states, typical causes and effects (pace tendencies, again) can be an analysis of "x is a sensation-of-red". But since those are the only sorts of relations allowed in Smart-type analyses, no Smart-style analysis is adequate.

12. The Identification of Sensations-Of-Red With Physical Tendencies.

We ask whether sensations-of-red, and their distinctive property, being a sensation-of-red, can be identified with physical states and a physical property. There are three principal candidates for the identification. First, being a sensation-of-red is identical with a causal tendency. Second, being a sensation-of-red is identical with a belief, or something belief-like. Third, being a sensation-of-red is identical with some physical property, the nature of which we do not in 1975 know, but which science can be expected ultimately to reveal. In this section, we will discuss tendencies.

If known differences in actual effects are not enough to account for all of what we know about our sensations-of-red, perhaps differences in their tendencies, or potentialities, or powers, to produce those effects are. The enterprise we are now engaged in, is to find some acceptable physicalist property for the property, being a sensation-of-red, to be identical with. And while actual effects and their similarities and differences do not have the pattern of similarities and differences that they would have to have were they to be identical with being a sensation-of-red, being a sensation-of-green, etc., it might be thought that tendencies to produce effects do.

For one thing, different causally relevant properties will at least tend to produce different effects, i.e. will produce them in some possible situation. So there is a prima facie case for thinking that the differences between the tendencies to produce causal effects line up one-one with the differences between sensations-

of-red, sensations-of-green, etc. For another thing, what we know about our states has the convenient feature of a certain referential opacity: we can know that Fa, it be the case that a = b, and yet fail to know that Fb. So, it might be said, there is nothing in the (undoubted) fact that we do not often enough know that our states have any causal tendencies, to prevent both its being the case that we know that our state has the property of being a sensation-of-red, and also its being the case that the property being a sensation-of-red is identical with a given causal tendency.

Such a move would be gratuitous without some indication of <u>which</u> tendency (i.e. the tendency to produce <u>which</u> effects) is the candidate for identification with being a sensation-of-red. Clearly, sensations-of-red produce very many different effects depending on many factors. For every occurrence of the state of seeing a red object, there will be a matrix of possible causal outputs from that state determined by the rest of the causal input at the time (e.g. the rest of what is being looked at), and the rest of the internal state of the person at the time (their other mental states, beliefs, emotions, their memory banks). The causal outputs do not necessarily constitute <u>actions</u>, the states in question might just issue in long or short term changes in the total internal state, emotions, beliefs, goals, memory, abilities e.g. sorting abilities, or just physical states. It is rather like a complex Turing machine table, and it is convenient to think of the possible outputs this way. Given a different sort of sensation-of-red, e.g. the state occurring when we allegedly have a red after-image, there will at a given time be a different matrix

of outputs, a different machine table.[10]

We are looking for a tendency with which to identify that which is common to various sensations-of-red. The machine table gives us plenty of tendencies e.g. the tendency to cause the report "I see something red" if asked and in a co-operative mood, the tendency to halt before crossing the road if in a certain sort of surrounding visual and belief state, and so on. For every entry on the machine table "If S(stimulus) and I (internal state), then R (sensation of red) causes E (effect)", we have a tendency: the tendency for E to occur if S and I. But surely the point about such tendencies is that they are <u>distinct</u> from one another, and hence if we identify any one with the property of being a sensation-of-red, we ought on a principle of parity identify the others, and this would lead to the intolerable situation that the distinct tendencies (a) to cause the report "I see something red" if asked and in a co-operative mood, and (b) to halt before crossing the road if in a certain sort of visual state, would have to be identified.

We need to find a tendency which is present in all cases of sensations-of-red for a particular person (for if it is to be identical with the property being a sensation-of-red, then that is present in all cases of sensations-of-red). Perhpas the argument of the previous paragraph would lead us to want to identify the <u>conjunctive</u> tendency, thought of as the tendency " to E_1 if S_1 and I_1 and to E_2 if S_2 and I_2 and ...", with being a sensation-

---

[10] Notice here that such machine tables bear a strong resemblance to Ramsey Sentences.

of-red, where the conjuncts in the above are obtained by listing the whole machine table. But the machine table for what? Different sensations-of-red e.g. seeing-a-red-object states and allegedly-having-a-red-afterimage states, have different, and to some degree incompatible machine tables. Yet we are after something in common to all sensations-of-red.

The example of pain fits this model of some kind of constant tendency present in all instances. It is not difficult to believe that there is a tendency whenever we have a pain to make some sort of characteristic expression of it. Even if the machine table for pain is quite complicated, there appears to be some loosely circumscribed class of effects the tendency to which we might claim is always present, even if there are tendencies to produce other effects also present in various situations.

On this model, we might argue about visual states thus: that there is always present the tendency to respond as if we were seeing something red, unless we are prevented from forming the belief that we really do see something red, we will for that belief. The natural way with afterimages is to believe that they are real.[11] So here we have something common: the tendency to believe that there is something red in front of the eyes etc. Furthermore, there are other causal effects that characteristically accompany the belief that there is something red in front of the eyes e.g. memory

---

[11] Children apparently often reach out to touch their afterimages. The response to the afterimage state as an afterimage state is learned.

effects, the creation of the ability to sort the (imagined) object into various similarity classes in respect of colour etc. So we can say that there is always present whenever a sensation-of-red occurs, the tendency to produce the belief that there is a red object in front of the eyes, the tendency to produce the memory that there was a red object in front of the eyes, the tendency to produce the ability to sort the (imagined) object into similarity and difference classes, discriminate it in a field of daisies etc.

This is very tempting. I cannot see any identification of the property, being a sensation-of-red with tendencies working unless this does. In fact, our discussion of beliefs in the next chapter will lead us to precisely the same spot. For that reason, I wish to defer the bulk of my discussion until then. Nevertheless, there is a point which can be made which is fairly damaging to this proposal. The point is, that anybody who believes that the tendency to believe that there is a red object in front of the eyes must still be present when we allegedly have an afterimage, is committed to holding that this tendency cannot be abolished by learning. One of the ways in which learning modifies human behaviour is to abolish certain behavioural tendencies and to supplant them with others (i.e. modify the machine table). It is certainly an unusual picture of learning to think that all that happens through all our lives is that we accumulate behaviour dispositions, in many cases inconsistent ones, without the extinction of any.

The extinction of a behavioural tendency would seem precisely to be the case in coming to grasp the

difference between seeing something red and having a red afterimage, or an hallucination of something red. We learn that there is nothing red there, that responding as if to a red object is inappropriate.[12] Anyone who insists that even after a great deal of suitable conditioning, the tendency to believe that there is a red object must still remain is surely guilty of rank a priorism. Or worse, for surely there are ample actual cases where it is false that we are aware that we have such tendencies. If we ask whether the learning process designed to extinguish the tendency to believe there is a red object in such situations has been successful, what better evidence in the present state of science for the conclusion that it has been successful, than that when we try to find it, it is no longer there. If you insist that there must nevertheless be such tendencies albeit unconscious, this would cause one to wonder what would convince you that you are wrong.

Another argument for the same conclusion is this: that the identification of sensations-of-red with tendencies to believe in red objects would seem to prevent our coming to realise that solipsism were true if it were. I do not mean of course that solipsism is true, but that it might have been true and we continue to have sensations-of-red, and we come to realise this. One way this might come about, is that you are sitting here in this room and all of a sudden things go wild, with shifting colours, no stable shapes, or recognisable shapes of physical objects, no stable sounds, and even your body

---

[12] There is of course no logical necessity that the appropriate responses be learned in any way, as we have pointed out before, and so no reason why someone should not be born with no tendency to believe that their afterimages are real.

disappears. You might well come to the conclusion after a time that you were the last person left in existence and that even physical space had gone, and it might be true. You might after some time come successfully to conquer the tendency to believe that your sensations of red are caused by red objects. Another way, not strictly a case of solipsism is this: Dr. Doom captures us all, puts us into the brain machine and gives us beautiful experiences which we believe are real. After a while, the machine temporarily malfunctions, and kills Dr. Doom and everybody else but you. It then rights itself and goes on producing dreams for you alone, the only mind left in the universe. Dr. Doom was aware of this possibility, however, and so programmed the machine that if it occurred, it would from time to time give out clues (dropped hints, memory jogging about Dr. Doom's takeover, even the occasional spoken or written sentences). Eventually you deduce what has happened, and why your experiences were curiously different from what they were before. You deduce that your sensations-of-red (you learned what they were before !) are not caused by a red object , that perhaps there are no red objects in your vicinity (if only you could get out of the machine !). After a time you come to lose the tendency to believe that there are red objects in front of your eyes.

Somebody might reply to the first case that if solipsism were true, then sensations-of-red would not be identical with anything physical, and perhaps not be identical with tendencies either, but that does not prevent their being (contingently) identical in this world. The second case, however, would seem to be consistent with human physiology being as it is now, and thus sensations-of-red continuing to be identical with whatever physical items

they are identical with in our real world. The case does seem to show that one could continue to have sensations-of-red and yet lose the tendency to believe that there are red objects in front of the eyes. If someone continues to insist that in such a case the tendencies would still be present, just gone underground. I do not know how to refute them but I cannot think of any reason for thinking that what they say is true.

I conclude, then, that sensations-of-red are not tendencies or dispositions. There is, however, a very considerable amount of argument about whether a theory of perception called Direct Realism is true. A consequence of Direct Realism (not necessarily noticed by its proposers) is that sensations-of-red are repressed tendencies to believe that there is a red object in front of the eyes. In the next chapter, we will be discussing this theory, particularly the version proposed by David Armstrong. I will be arguing two things: that sensations-of-red are not what the theory must say they are, and that a different theory of perception, best described as a version of Representative Realism, is true.

CHAPTER TEN. PERCEPTION

1. Theories of Perception.

Reflection on the fact that human beings are subject to illusions and hallucinations has led some people to suppose that perception of the external,[1] physical world is always mediated by some sort of visual experiences, sense data, or raw feels. A sense datum was held to be caused by the physical thing of which it is the sense datum, and to this extent may be said to represent the physical thing.[2] On the basis of our awareness of the sense datum, we make an inference about the nature of the external world causing the sense datum. The inference gives rise to our ordinary beliefs about the external world, and they in turn might or might not be true. When they are not, some sort of perceptual aberration, perhaps an illusion, has occurred. The view is realist about the physical world, and so it is called Representative Realism.

This account of perception has received extensive criticism. It is by now quite common to argue that the theory is defective in at least two respects. First, that on the theory that we can have no good grounds for believing in the external world at all. The alleged inference that takes place can never be a good one.

---

[1] "external" henceforth means "distinct from human beings and their parts".

[2] A more complicated (and less plausible) theory might also hold that the sense datum represents the physical thing in that it is like that thing. We will not discuss such theories.

Second, that the theory seems false to the manifest facts of perception in that we do not, at least consciously, make any such inference.

These difficulties have led other philosophers to propose an alternative account. Being realists concerning the external world, and so unwilling to accept the lack of knowledge of the external world to which Representative Realism seems to doom us, they have wanted to say that our perceptual awareness of the world is direct, that is to say, unmediated by any sense datum of which we should otherwise need to have prior awareness. We call this view Direct Realism, but it is a group of views varying in the account of what directness comes to.

2. Direct Awareness.

We have so far in this thesis eschewed the locution "aware of" in favour of "aware that" (except where it seemed clear that no harm has been done). Also, we have been using "aware that" and "aware" to mean "know that" and "know". Continuing this practice, we distinguish four senses of "x is directly aware that p". [3] In all four senses, x knows that p. The first two senses might be called "ontological" senses in that they propose a mechanism (of sorts) for the knowledge that p to be direct. The first: "x is acquainted with the fact that p". [4] Acquaintance might be thought of as like

---

[3] These definitions derive in part from Cornman 1972. On acquaintance, see Russell 1959. On the fourth sense of "directly aware", see e.g. Malcolm 1963.

[4] An ontology of facts or states of affairs is apparently forced upon us in order to generalise over

a reaching out of the cognitive faculty of the mind to include the state of affairs, with no obstruction between the knowledge and the state of affairs because there is nothing between the knowledge and the state of affairs.

The second: x is directly aware that p iff nothing mental lies causally between the state of affairs that p, and x's knowledge that p. The state of affairs that p is causally responsible for x's knowing that p, but the causal chain does not include anything mental e.g. a sense datum, to throw up a "veil in front of our awareness".

The other two sensesmight be called epistemological senses. The third sense: x is directly aware that p iff x does not make any inference in coming to know that p; if x's knowledge that p is noninferential. The fourth sense: x is directly aware that p iff x logically cannot be mistaken about whether or not p.

The fourth sense is obviously not a sense in which we are directly aware that there is a piece of red brick in front of our eyes. Our perceptual mistakes about the presence or absence of pieces of red brick are well known. As for the first sense, acquaintance is barely intelligible. I mention it, though, because I believe that some Direct Realists unconsciously rely on it even though their doctrines do not contain it as a part. I will say more of this later.

---

what follows the "that" in "x is aware that ...". It often happens in philosophy that locutions describing particular situations present no philosophical problem, but that the generalisations forced on us by the temptations to theorise create their own problems.

Armstrong and Pitcher[5] are Direct Realists in both the second and third senses. There is a reason why someone might hold both senses together. To see this, remember that inference for Armstrong is a causal process: to know that p and to infer that q from p, is to have one's knowledge that q caused in an appropriate way by one's knowledge that p.

Now if, as was traditionally held, sense-data are necessarily conscious items, then if we allow a mental item like a sense datum to intervene causally between the state of affairs that p and the knowledge that p, we are certainly running close to the idea that there is an inference from the knowledge of the sense datum to the knowledge that p. Conversely, if nothing mental lies causally between the state of affairs that p and the knowledge that p, then the knowledge that p does not involve the making of any inference.

The Direct Realism I will be discussing is the conjunction of the two senses, two and three. I will be denying that the two senses necessarily go together, and claiming that Direct Realism in sense two is false, that Direct Realism in sense three is true, and sketching a view which deserves better to be called Representative Realism than Direct Realism. (RR and DR hereafter.)

3. Advantages of Direct Realism.

First, let me say what a good theory DR is. Its principal merits lie in its economy, and its ability to resist most of the traditional arguments for RR (and so against DR).

---

[5] e.g. Armstrong 1961, 1968. Pitcher 1971.

In denying sense data, it denies a class of mental items that are at least prima facie problematic for the physicalist. If affirming that there is such a thing as knowledge of the external world, it affirms something which RR also affirms. The causal mechanism of DR is simpler than that of RR: it has one less kind of item. Furthermore, prima facie it escapes the two problems for RR that we have already mentioned: it claims that there is no veil of sense data to break through to obtain knowledge of the external world; and it makes no (patently false) claim that we are constantly making inferences to the external world, or even that we ever do.

It resists easily such traditional arguments for RR as the Time Gap argument, the Argument from the Scientifically Established Causal Chain, and (certain versions of) the Argument from Illusion. I will not discuss these at great length, but I will sketch them to show why they do not establish RR, by showing that DR is (easily seen to be) consistent with them.

The Time Gap argument points to the fact that because causal signals (e.g. light) have a finite upper limit, the onset and ceasing to be of the knowledge that p invariably lags behind the onset of the state of affairs that p and the ceasing-to-be of the state of affairs that p. In some cases, e.g. when seeing stars, the time gap is very great. Still, there is something that we see at the time, and so, goes the argument, this something must be a sense datum. This argument is defective because nothing in DR prevents the knowledge that p occurring later then the state of affairs that p. DR can also agree that there must be something that we see at the time.

That is what seeing the star at the time amounts to: having one's beliefs about the star caused in an appropriate manner.

The Causal Chain argument is neatly summed up by Pitcher:

> ... since the awareness we have in sense perception comes at the end of a causal chain, the objects we are thus aware of cannot be identical with whatever it is in the "external world" that figures in an early stage of that causal chain.[6]

Both this argument and the previous one seem to get their force from the misconception that direct awareness of the external world must be acquaintance with it, i.e. cannot be separated from it. But if we take awareness to be direct in sense two, then there is nothing to prevent our knowledge of the world coming at the end of a causal chain that begins with the objects of which we have knowledge and which can be temporally separated from that knowledge. And so there is no need for the introduction of sense data. Indeed it is this separation that makes certain kinds of illusion possible. Knowledge is a belief of some sort, and if the state of affairs and the belief that it obtains are separate, then there is no reason why the latter cannot occur without the former i.e. no reason why the belief be false. But that is precisely the condition for one sort of perceptual illusion: that our beliefs acquired by perceptual means be mistaken. Thus versions of the Argument from Illusion which point to the existence of illusions, i.e. perceptual mistakes, in order to demonstrate sense data cannot succeed. DR easily accomodates perceptual mistakes.

---

[6] Pitcher 1971 p.44.

Similarly, facts about perceptual relativity are easily accomodated. The fact that people perceive the one object from different viewpoints does not force us to accept private sensory objects to correspond to the varying viewpoints: people just acquire different beliefs about the object and surrounding conditions.

So DR is a very strong theory. There is one argument which it cannot resist, however. It is a version of the Argument from Hallucinations, although in the context of this book it might equally as well be termed the Argument from Afterimages. This argument is one of the reasons why this chapter is included in the book, because it is also another part of our more general argument against materialism. I will try to show that in perception more mental items occur and are causally operative than mere beliefs, in particular that items quite like sense data occur. I will argue that these items are neither beliefs, nor suppressed tendencies to believe, nor any such belief-like item.

4. Is the Property, Being a Sensation-Of-Red, Belief-Like ?

Sometimes when we (allegedly, pace Rorty) hallucinate, as when we allegedly afterimage, we are in a state which we can know quite a lot about: similarities, differences, causes, effects and so on. Furthermore, if the argument of this book has been correct so far, we can know that our state has certain properties, which we have called "being a sensation-of-red," "being a sensation-of-green," etc. Now these same properties, it has been argued, are typically present in our states, and can be known to be present, when we see something red in normal conditions, see something green in normal

conditions, etc.

The first thing I want to say is that it is the presence of these properties in our states which, under normal conditions, is a necessary condition for our believing that there is something red in front of our eyes[7] (someone who would identify such properties with beliefs should hardly dispute this). Hence, it is a necessary condition, in normal circumstances, for our seeing something red. How can we establish this beyond noting that it is obviously true? The property in question, being a sensation-of-red, is what is in common to allegedly-having-a-red-afterimage states, allegedly-hallucinating-red states, and seeing-a-red-object states. Take away that property from our state, replace it with the property common to allegedly-having-a-green-after-image states, allegedly-hallucinating-green states, and seeing-a-green-object states, leaving everything else the same, and you will find that in normal conditions humans believe that they are confronted by a green object.

In fact, leaving aside the causal genesis of our state, every other conscious feature of our state when we see something can be duplicated in an hallucination, including the "attendant" beliefs and causal outputs. We can have a full blown hallucination, believe that "it is real" and act accordingly.

Now if Direct Realism is true and the only mental items necessarily present in perception are beliefs, then sensations-of-red, which I have claimed are necessary for the perception by humans of red things, must be beliefs or at least in the same conceptual category as

---

[7] To select, for convenience, the property of being a sensation-of-red.

beliefs. Why I say that sensations-of-red are worth being called mental, though I have not given a definition of "mental", is that they are states of humans which they can come to be aware that they have, and apparently without inference.[8] The alternative for the Direct Realist position would be to deny that such states and their special properties are mental, and so to allow that it is consistent with DR that they occur in perception. But that would be a cheap victory. One might just as well redefine "mental" so as to disallow sense data from being mental, and so accomodate RR under DR in sense two. In any case, it is certainly in the spirit of DR either to deny the existence of features like being a sensation-of-red or identify them with belief-like items.

Is being a sensation-of-red belief-like? Surely it does not involve a conscious belief about the external world, for we can have full-blown hallucinations and afterimages with this feature and yet have no introspective knowledge of any beliefs about the external world.

Perhaps we might have the belief, but unconsciously? Perhaps the property, being a sensation-of-red, is identical with the property, being a belief that there is a red object in front of my eyes, but we are not aware of the property <u>as</u> a belief? That is to say, perhaps we just know that we have the property, the property is identical with the belief-property, but we fail to know that we have the belief-property.

---

[8] To call them "mental" here is not to deny that they might also be physical. Nor do I wish to deny that inference might be necessary for a particular person to come initially to recognise the existence of such states, e.g. a Direct Realist philosopher.

Aside from the possibility that we know that our state has the property, surely the principal evidence for whether we have the belief would be whether we act on it -whether it is causally operative. Beliefs have a habit of changing people's responses to the world. But it is precisely cases where we are not fooled by hallucinations and afterimages, that we do not act as if there were really red objects in front of us.

Neither Armstrong nor Pitcher make this move. Rather, they say that what is present in hallucinating and afterimaging are things like suppressed inclinations to believe, unconscious tendencies to believe. With this we arrive at the point we arrived at the end of the last chapter. Notice that the Direct Realist will have to identify the property being a sensation-of-red with some belief-property e.g. the property of being a tendency to believe that there is a red object in front of me. If that identification is not carried through, then sensations-of-red are not wholly belief-like. They have a feature which is not just the property of being a tendency to believe, and we have already concluded that this feature is causally vital in normal perception.

Since as we have already pointed out there are no such known tendencies in certain cases where we have a sensation-of-red, the Direct Realist would have to say that the tendencies are unconscious. The opacity of "know", however, allows us to accomodate this fact with the proposed identification. But why should we make the identification? After all, how would we refute the identification? Surely the beginnings of a refutation, is that when we look for the candidate to identify our troublesome property with, we do not find it.

6. Pitcher's Reasons for Claiming That a Belief-Tendency is Always Present in Illusion.

Surely the worst of reasons for making the identification are Pitcher's. Let us look at them briefly. He considers three sorts of cases of perception, which he calls First Cases, Middle Cases and Last Cases. First cases are normal, standard cases, in which a person

> causally-receives in that way (the normal causal way - C.M.) the belief, which he does not question, that there is indeed an x at u.[9]

Middle cases are cases where there is no disputing the existence of a tendency to believe that the world is a certain way

> The driver, we may say, half-believes that there is a pool of water on the road ahead, or, as I shall prefer to put it, that he is inclined, or has an inclination, to believe that there is such a pool.[10]

And Last cases:

> ... are marked by the fact that although it looks to Q as though there is an x at u, Q nevertheless does not causally-receive the perceptual belief that there is an x at u - on the contrary, he acquires the firm belief that there certainly is not an x at u.[11]

What are <u>perceptual</u> beliefs for Pitcher? I think that they are a red herring for our purposes, but I will sketch Pitcher's account in order to satisfy the reader of this.

On p.70, Pitcher defines perceptual beliefs in terms of when they standardly occur.

---

[9] Pitcher 1971 p.86.

[10] ibid., p.92.

[11] ibid., p.92.

> I shall can any belief ... that is acquired by using one's sense organs in standard ways - a perceptual belief.

Later, he recognises that he will want to say that such beliefs are present when the standard causal genesis is absent. So he says

> ... by a perceptual belief that there is an x at u I mean one that a person has when, in first cases, it looks (in the phenomenal sense) to him as though there is an x at u.[12]

> ... what I shall call a phenomenal sense of "looks" - i.e. the sense of "looks" involved when it can be said of a person who is looking, under perfectly normal conditions, at a pencil lying on his desk, "It looks to him as though there is a pencil lying on his desk".[13]

His account of this last "looks" locution is, however,

> Q causally receives, by means of using his eyes in the standard visual way, the (perceptual) belief that there is an x at u.[14]

Round in a circle ! So we must either take "perceptual" as primitive, which is a good idea neither for physicalism nor for Direct Realism, or we must return to the original idea that to say that they are perceptual is to indicate standard but not universal conditions in which they occur, but not to indicate anything else special about them.

Now the hard cases for DR are Last cases, and, I am claiming, cases of hallucination and afterimaging where we are not fooled. Nevertheless, Pitcher wants

---

[12] ibid., p.90.

[13] ibid., p.86.

[14] ibid., p.90.

to say of them that

> the perceiver may plausibly be said to causally - receive an inclination to believe that there is an x at u, but ... it is an inclination that for some reason or other he resists or overcomes, one that he quashes or strongly suppresses, so that it is an attenuated inclination. I shall say that he causally receives a suppressed inclination to have a perceptual belief that there is an x at u. [15]

Now as we have already noted, there are cases where such an inclination is "suppressed" to the point of our never being conscious of it. Why should we believe the inclinations are always there ? Here is Pitcher's reason :

> This kind of inclination is to be regarded as a theoretical perceptual state posited by our (new) theory of perception in order to account for certain difficult cases. [16]

So if you look for them and cannot find them, you should still believe they are there (postulate them !), otherwise you will be refuted !

Perhaps this is a little unfair to Pitcher. One might be inclined to postulate such unobservables if one thought that all alternative theories to one's own had insuperable objections to them. However, I shall argue later that the usual objections to Representative Realism can be answered in the version of it that I am proposing.

One point worth making is that the line I have been arguing is somewhat stronger than necessary to establish my position (though I cannot think of a convincing way of arguing a weaker line). For if someone manages to establish against me that even in full blown

---

[15] ibid., pp. 92-3.

[16] ibid., p. 93.

cases of hallucination and afterimaging where we are not for a minute fooled, there is nevertheless still present an unconscious tendency to believe that there is a red object in front of the eyes, then they still have to show that the property of being this belief is identical with the property of being a sensation-of-red. If they do not do that, then DR is undermined, and the broader defence of physicalism concerned to de-fuse sensations-of-red by identifying them with beliefs, fails.

There is another argument for the conclusion that there is more to perception than tendencies to believe, however. It is another argument in addition to the two already given, (namely the manifest absence of the tendencies in some cases and, in the previous chapter, the possibility of learning being successful to the point of extinguishing the tendencies). The argument arises out of an attempt to refute Representative Realism with the Case of the Speckled Hen. It is an argument from the complexity of our perceptual states.

7. The Speckled Hen.

The discussion so far suggests the following picture of perception: there occur in perception two sorts of mental items, one sort like sense data in that they are not beliefs and not belief-like, and in that they are typically causally sufficient in normal situations to give rise to the other sort of item, namely a belief about the external world. We have argued that the first sort of item includes such things as sensations-of-red, sensations-of-green, and we will give them a general name: perceptual-sensations. The sufficiency in question is undoubtedly causal sufficiency: perceptual-sensations typically (but not always) cause in perception our knowledge

(beliefs) about the external world. The trick is to be able to say this without committing oneself to the admittedly false view that we are constantly making an inference from our perceptual-sensations to the world.

We have already argued that mental items are corrigible, in the senses that one can be in a mental state and fail to believe it, and fail to be in the mental state that one believes one is in. In fact, something stronger is true: that it takes a special sort of sophistication to know that one has mental states in perception, and a certain effort to come to know it on a given occasion. (Anthony Quinton makes this point very effectively.[17])

In normal cases of perception we are attending to what if going on in the world, typically with no thought for ourselves. Indeed it is precisely this phenomenon which makes it implausible that we make any inference in perception.

The phenomena of attention are interesting in various ways. We can pay attention to our sensations[18], or pay attention to physical objects out there and not notice our sensations. When driving a car, sometimes we are conscious of the trees flying past on the side of the road, and sometimes not. But when we are not, this does not always seem to be a case of just not having any visual sensations at all. This is for several reasons: we can come to be aware of the sensations by a small act of will, and it is not like bringing the sensation into existence, but more like becoming more conscious of

---

[17] Quinton 1955.

[18] Perceptual-sensations, that is.

something already there. Again, there is the phenomenon of being dimly conscious of those sensations, and then becoming more fully conscious. Thus there appear to be degrees of consciousness of perceptual-sensations and their features, and if the sensations themselves stay constant, ("we should have been more aware of them, but were not"), then the sensations and our awareness (knowledge) of them are distinct.

Or again, something plays the same sort of causal role in our behaviour, as that which we at other times know to be a perceptual-sensation. We automatically make avoidance behaviour when driving a car, sometimes without realising that there was something we were avoiding, or without realising anything about our states. It is well-known, too, that a very brief exposure to visual stimuli need not evoke any conscious response in us, and yet certainly register on us. Subliminal advertising is an example of this. There can even be the phenomenon of remembering something much later which we were at no time conscious of (and which may have been either a mere perceptual-sensation, or a full-blown visual experience), because our attention was on something else. Now in cases where things happen too fast to trigger off conscious awareness, people can sometimes be trained to come to be aware of having a sensation. Furthermore, such people can come to be trained to be aware that it was the sensation that was causally operative in a certain way, whereas before all that was known was that <u>something</u> (unconscious) was causally operative in the same way. This last point is a particularly strong one, I think. A person can be exposed to stimuli, register nothing consciously, but have their behaviour modified. At a later occasion, the

person can learn to recognise that they characteristically have a certain mental state caused in an identical stimulus situation, and that it is the mental state which produces identical behaviour modification. This gives a reason for dignifying the underlying conscious cause as "mental"- namely that it occupies a similar causal position to another state which is mental - even if we are unable to accept that we learn to recognise it for what it is, namely a perceptual-sensation.

But if, often enough, when we are perceiving, we do not have any beliefs about our mental states, then it cannot be that we infer our beliefs about the external world from our beliefs about our mental states. A mark of this absence of any actual inference is that we are not conscious of having any beliefs (often enough) about our mental states. Another mark of whether an inference actually takes place, is that we offer as reasons those which the inference comes from. Certainly most people, even philosophically sophisticated people, would be non-plussed if asked for their reasons for saying that they see a horse, when faced by a large brown Clydesdale at ten paces in broad daylight.

But if we agree that we are not aware of our perceptual-sensations in the normal perceptual situation, then it is easy to agree that we do not make an inference from the sensations to the external world. The causal role of the sensation is accommodated by saying that the sensation typically directly (i.e. with no mental causal intermediary) causes the belief in the external world. There is no prior belief about the sensation to be a causal intermediary, therefore we do not have to say that an inference has been made.

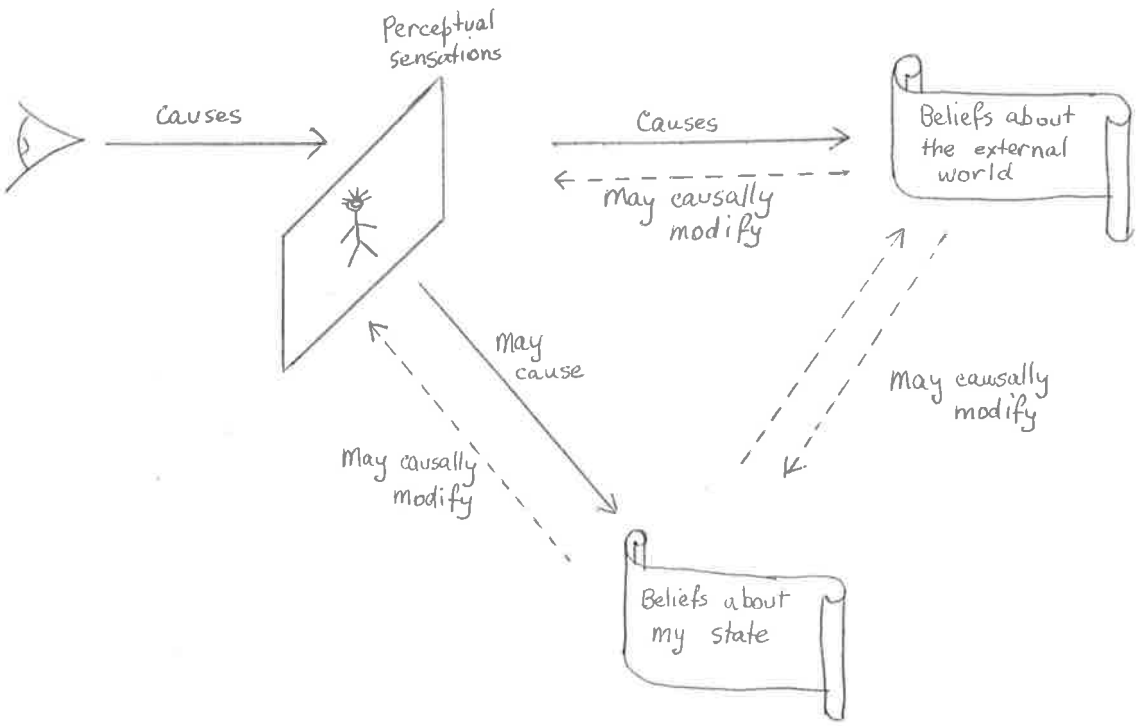

Fig 2. A Theory of Perception.

What then is the place of our occasional knowledge of our perceptual-sensations? When it occurs, does it causally intervene and set up an inference to the external world that was not there before? Not necessarily, and not usually. Armstrong's model for the usual sort of introspective awareness, likens it to a scanning mechanism, scanning other of our states. The scanning seems to me to be most economically described as a causal process: we are in state S, we come to know that we are in state S by having state S cause our belief that we are in state S. There is no reason on this (plausible) model why the knowledge that we are in S should causally affect the belief that the external world is a certain way (although it might effect it in some cases). Figure 2 is a picture of the sort of thing I have in mind.

Such a picture of perception is sufficiently like classical Representative Realism to be worth the name, differing only in separating sense data from their consciousness, and eliminating inference to the external world (even though in doing so it becomes Direct Realism in sense three of "direct"). It resists the second of the two criticisms of RR given above in section one (namely that the inference to the external world usually does not occur). Whether it can resist the first (i.e. no reason to believe in the external world) we will see later. But Armstrong gives a third criticism of RR.[19]

Imagine that we are looking at a speckled hen. When the hen has a lot of speckles on it, normally we

---

[19] And a fourth, the charge of "unproven". We will look at this later although our argument in the previous section is the bones of a proof.

cannot tell at a glance how many there are. But if perception consists in the having of sense data then we must say either that our sense datum of the hen has an indeterminate number of speckles, or that it has a determinate number but we do not know how many. The first alternative requires us to believe that there is an entity with some speckles but no particular number (or, requires us to believe that there is an entity whose properties are indeterminate). The second alternative requires that sense data have properties of which we are unaware.

Armstrong's version of the argument suffers from a defect which I think can be fixed up. He assumes, apparently uncritically, that we will say that the sense datum of the hen is speckled, as well as (presumably) the hen itself. This is surely questionable.

Still, sense datum theorists have wanted to say that we are aware of the external world in virtue of being aware of sense data, and so presumably they must hold that there is something about the sense datum of the hen, some feature of it, which is the basis of the judgement that the hen has speckles. We might well call it "speckles*". Furthermore, there must be something about speckles* which could lead us to say in (simple) cases that the hen had two speckles rather than three. Something like number; we might call it "number*". The speckled hen argument will then show that either some sense data are indeterminate with respect to the number* of speckles* that they have, or none are and we are unaware of some of the features of some sense data.

In reply to this dilemma, one reason people have had for introducing sense data is the alleged "indubitable

core" of perception: in some cases where we are in doubt about the nature of the physical objects that we see, still there is something that we see, something whose properties we cannot be in doubt about (and so that something must be internal, mental); thus, incorrigible sense data. If you believe in sense data for this sort of reason, then indeed the second horn of the dilemma is unacceptable.

Having some number* of speckles*, but no particular number*, is not so obviously intolerable. In my view, however, the number* is really the number, and the idea of there being some number of things without there being any particular number, is unacceptable. So I wish to avoid grasping the first horn. The second horn, however, provides no problem to someone who is prepared to distinguish between our perceptual-sensations and awareness (knowledge) concerning them. (If one does not allow that sense data logically could be unconscious, reflect that our description of the theory being argued for did not use the word "sense datum".) In fact, as has already been argued in Chapter Five, the speckled hen example provides an argument for just this distinction to anyone who is already prepared to accept that in perception there are perceptual-sensations in addition to belief-like items: notice that your sensation of the hen has a definite feature, namely a certain number of speckles (or speckles*), but that you are not aware of how many. Needless to say, one ought not in the interests of consistency be prepared to believe in perceptual-sensations on the grounds of the incorrigibility of certain elements in perception (c.f. previous paragraph). But then this is not the reason given here for believing in

them.

We can give a similar analysis of another often-cited case: the three colour problem. It is well-known that there can be triples of objects, say squares of paper, which are very alike in colour, but which have the curious property that A and B are indistinguishable as to colour, B and C are also indistinguishable, but A and C when put together are distinguishably different. Transpose the whole case into afterimages or hallucinations, and the same can obviously happen. Do we have to concede indeterminateness as to colour of our three afterimages? We do not have to, and I am loath to admit that colour-properties are the sort of property that can be indeterminate in this way (i.e. that something can be some colour but no particular one). (This is not the same as saying, of course, that colour is determinate as to description.) It is much more reasonable to suppose that we cannot discriminate the relevant features of our states sufficiently finely to distinguish having-afterimage-A from having-afterimage-B. This fits, too, with the fact that people can learn to discriminate differences where they previously did not do so, even in afterimages, hallucinations, or in identical stimulus situations.

These considerations do not mean that we must always give second-horn-type solutions to similar problems about perception. Imagine standing up close to a book and being able easily to read the letters on the spine. Then imagine slowly moving away from the book until we can no longer read the letters on the spine. Now imagine that it was all an hallucination, and ask: at the end point, are the letters in our hallucination somehow

indeterminate, or do they perfectly determinately spell out the title of the book but we are unaware of this? The second horn is unreasonable; surely it is not the case that every time we fail to discriminate in perception there really was some feature of our sensation there to be discriminated (or to form the basis of a possible discrimination of the properties of the external world). It is easier to say that at the end of the continuum of cases, the relevant features of our sensation have become indeterminate as to letterhood. Indeterminateness as to number (with some exceptions) and indeterminateness as to colour (barring quantum theory troubles) are not so easy to accept, but indeterminateness as to letterhood is certainly not so difficult.

8. The Speckled Hen Turned Against Direct Realism.

I think that the Speckled Hen example can actually be turned against Direct Realist theories. This will be our fourth argument that being a sensation-of-red is not something belief-like. In fact, it is not a direct argument that this particular property is not belief-like, but, rather, depends on an argument that another property necessary to perception as we know it in a similar way to the way that being a sensation-of-red is, is not to be identified with anything belief-like.

Consider hallucinating a speckled hen. Now the thing about our state at the time is that it is a complex state: we can discriminate various properties of it, can know e.g. that some of them and not others are identical to certain properties of certain states had on other occasions, and so on. The state we are in can be similar in various important ways to the state we are in

when under normal conditions we see a speckled hen. It can be similar in one important respect, in that a mental item necessary for the normal case of perception can be present in the hallucinatory state. Now if Direct Realism is true and all that is mental and necessary for perception are beliefs and belief-like items (suppressed inclinations to believe), then that is all that is mental and present when we hallucinate the hen.

The hallucination-state is complex, therefore we should need either a complex belief, or a complex of beliefs.[20] Pitcher recognises this point, and speaks of the "richness" of our perceptual beliefs.[21] What I shall try to show is that our state has a feature, which we can come to know about, for which there is no corresponding belief.

Let us suppose first that we hallucinate the hen, and falsely believe that there really is a speckled hen in front of us. We believe that it has many speckles, but for any particular number x, it is false that we believe that it has x speckles. Now I claim that in this case, the beliefs and suppressed tendencies to believe which we have do not exhaust the content of what we can introspectively know. Suppose we set to work to count the speckles: 50. We come to believe that the hen has 50 speckles. But what seems clear is that the number of speckles that the hen had (i.e. the relevant feature of the hallucination state) need not change during the counting process. We can "keep an eye" on the hen while counting

---

[20] If part of the hen is red, part is green, for instance, we will at least need the belief that there is a (1) hen, which is (2) part red, and (3) part green.

[21] Pitcher 1971 pp.88-90.

to check that the number of speckles does not change. It really had 50 speckles all the time, though we did not know this at first, and later came to be aware of it. We came to have a belief about the precise number of speckles. Perhaps (though I am disputing it !) it is the case that a description of the set of belief-like items is sufficient to specify the content of what we introspectively believe after the counting. But it cannot be that the beliefs we had before the counting exhaust what we know. Our hallucination-state had a definite feature which was unchanging through the counting, and which we came to express by saying "50 speckles"; and that feature was something that we did not have any beliefs about before the counting. The key point is that it is grossly implausible to say even that I had a suppressed tendency before the counting (perhaps I "overcame" it?) to believe that the hen had 50 speckles.

We might say that before the counting I had a belief (no suppression) that the hen had a definite number of speckles. No doubt that is often true, but it will not distinguish between having an hallucination of a hen with 50 speckles, and with 49 speckles, and with 5 speckles.

The whole argument can be repeated with the modification that instead of mistakenly believing that there is a hen there, I know that I am hallucinating. In this case there are no beliefs about the external world, which are essential for me to ~~when~~ I come to believe that I am in a state which is an (alleged !) hallucination of a hen with 50 speckles. I recognise that the state has a certain property which we can denote by "50 speckles" and that this property is different from ones denoted by "49

speckles" and "5 speckles". For Direct Realism to be true, this property ought, presumably, to be identical with the suppressed and unconscious inclination to believe that there is something in front of me with 50 speckles. But I have argued that this property can stay constant during the counting process,[22] and that, before that process, it is implausible to think that we had any beliefs about the number of speckles that had anything to do with fifty in them.

This argument has been about "number sensations" not "colour sensations", but once we see its point, the generalisation is not too difficult. The fact is that learning to discriminate the features of our mental states is a learned art, and, especially when our mental states are complex ones, an art which the majority of us could hardly be said to master. This is just as true as it is true that learning to discriminate the features in the world is a learned art. And it is hardly reasonable to think that there will be in us latent beliefs about the world corresponding to all the features of the world - or, if it is an hallucination-state we are in, all the features of our state - that we can later with learning or an effort of will come to know are present.

Can we construct a parallel example using perceptual-sensations of colour? It is a little harder, for colour seems intuitively so immediately present. Perhaps the following will do: we hallucinate, without being fooled, a rather extensive patch of streaky olive

---

[22] Counting spots on an hallucination ! Why not? People can read the words on their eidetic memory images. Surely this would seem strange only to a philosopher.

grey. After a while of careful examination we come to realise that there was a small streak of lighter green, almost yellow, in the right hand corner. Do we really want to say that all along there was present the unconscious tendency to believe that there was something in front of us which was (partly) a lightish greeny-yellow? We had no idea that the greeny-yellow was there ! Suppose we had really seen a patch of streaky olive grey sky but not noticed the lighter green patch. Would we still have wanted to insist that we had a suppressed tendency to believe?

This, then, is the fourth argument for the conclusion that the properties, being a sensation-of-red, -of-green, etc., are not to be identified with unconscious properties of being a tendency to believe that the world is a certain way. To recapitulate, the four arguments are the Argument from the Possibility of Extinguishing Tendencies, the Argument from the Possibility of Coming to Know that Solipsism were True, the Argument from the Unreasonableness of Pitcher's Postulation, and, most recently, the Argument from the Complexity of Hallucinations. I conclude here that the relevant properties are not to be identified with belief-like items.

The rest of this chapter is taken up with answering objections to RR, in order to complete our discussion of the philosophy of perception.

9. Objections to RR.

Objection 1. If you do distinguish between your normal perceptual awareness of the external world (and say that it involves a perceptual-sensation) and your awareness of your sensation, why not say that the second

awareness also involves a sensation? Do we not have a regress, or at least otiose and arbitrary second order sensations? And if you allow that certain cases of awareness, the introspective awareness of our mental states, do not require sensations, i.e. are direct, then what reason is there to resort to them in the case of perceptual awareness of the external world ?

The answer is that there is a reason if there is a reason. If there is an argument for sensations being present in our perception of the external world, then that is the reason for resorting to them. The argument need not be an argument for the general conclusion that all cases of awareness are cases causally mediated by sensations. (Such an argument would land us with objection 1.) The reason for "arbitrarily" stopping before agreeing to second order sensations is Ockham's Razor. In the absence of some further argument (and I do not want a priori to rule the possibility out) for the existence of second order sensations, it is better not to postulate them.

An apparent consequence of this, is that the theory of awareness of the external world I am advancing is a contingent theory, and I accept this consequence. It is not part of my theory that all cases of awareness be like this, and furthermore it is not part of my theory that perception would have to be like this in all possible worlds. We will return to this point at the end of the section.

Objection 2. A Representative Realist would have no good reason to believe that there is a physical world.

Armstrong notes than an inference to the external world might not be so bad if the inference is thought of as a piece of postulation to explain the order in our sensations, but comments:

> For surely we are not prepared to degrade bodies into hypotheses? We want to say that our assurance of the existence of the physical world is far stronger than any assurance we could obtain by indirectly confirming a theory. If the Representative theory were true it would be proper to have a lurking doubt about the existence of the physical world. Yet such a doubt does not seem proper.[23]

In my view, Barry Maund has completely answered this objection to RR in his "The Epistemological Objection to the Representative Theory of Perception".[24] The answer that follows makes some of the points that Maund does, and some different ones as well.

The first point I want to make in reply is that DR is in no better position vis-a-vis the justification of our beliefs about the external world, than RR is. In both theories it is held that we know about the external world. In neither theory are our (true) beliefs about the world in any way self-validating. In neither theory do we typically have reasons for these beliefs. What makes the beliefs into knowledge must be, on either theory, some causal story along the lines given by Armstrong. If either theory can agree that we do know in perception facts about the external world, then the objection that we have no reason to believe in it must amount, as Maund says, to the objection that on RR we cannot <u>know that we know</u> that there is an external world.

---

[23] Armstrong 1961 p.30.

[24] Maund 1974.

But how is DR any better off? If there is no reason to think that our beliefs about the world are true, why should the DR theorist believe that his or her beliefs about the world constitute knowledge about the world? Because there is a causal story to tell? But he can only know that story true if he knows some things about the world. Because, he holds, like Armstrong, that we must start somewhere by holding some beliefs of which we are certain ~~as~~ to be true until reason to disbelieve them is given?[25] But why should not RR do the same?

I suspect that the objection gets some of its force from the taint of the Acquaintance theory of perception that DR has. It is sometimes thought that RR introduces a veil between the mind and the world, which is absent in DR. But there is a veil of sorts in DR too, in that the knowledge of external objects is not right up next to the object: they are separated both in time and space. No mental object lies between them for DR, but then other objects, physical ones, do.

There is this much to Armstrong's idea: you do have to start somewhere. Some contingent propositions will have to be accepted without justification, or else there is regress. We have chosen at earlier points in this book to accept that we know some things about our mental states. It is not so bad to take a small further step and accept that we know some things about the world too. However, I do claim that in some sense we can justify this latter claim. For what better explanation could there be for the order and regularity among our mental events,

---

[25] e.g. Armstrong 1968 p.205. See also Maund 1974 pp.4-5.

than that many of them are the effects of a relatively stable set of causes occurring outside the stream of our mental states?

Maund argues that if we can assign some prior likelihood to this hypothesis, then we can produce ample confirmation tending to raise this likelihood to certainty. The prior likelihood, it seems to me, is found in the hypothesis being the best explanation.

Are there any alternative explanatory hypotheses for the stability of our experiences? There are three, I think. (1) no explanation for the regularities in our mental life need be looked for, (2) the explanation is to be found within our mental stream itself, (3) the explanation lies outside, in that it is not identical with any mental item of our own, but nor is it identical with any part of a physical world.

(1) can be dismissed as overly sceptical. The stability in our day-to-day experiences (e.g. experiences as of same colour curtains in my room, same sequence of light (day) and dark (night)) is too much not to justify intellectual curiosity as to its causes. If we grant this, then we can proceed to (2): such causes as there are of our mental states, are others of our mental states.

With (2), as with (3), we are in fact facing the old problems of the possibility of solipsism (or phenomenalism), and how to refute them. I do not think that such possibilities can be dismissed by an argument showing them incoherent. It seems to me that solipsism is possible, consistent with my mental states having been what they were up to now (excepting for those descriptions of them which logically relate them to external objects e.g. "knows", "is the effect of", etc.). I have tried

earlier to sketch solipsistic and near-solipsistic situations.

We can agree, with Freud, that some of the causes of my mental states lie in the unconscious, to add to the rather small store of mental states which we know to have a mental cause. But these do not begin to come to grips with the vast welter of highly structured perceptual "information" that we receive. Why, for instance, on "fine days", does the "sky" continue to look blue? And why when it does not look blue, does it not? Large masses of very regular perceptual-sensations and perceptual beliefs should just have to be accounted as inexplicable if we opt for (2).

One point needs to be brought out here. Notice that in describing the sensation in question, I have employed descriptions which tie them to external states of affairs: "experiences as of same colour curtains", "fine days", "sky", etc. It should not be thought that there is any logical necessity in doing so. There is no doubt that we do describe many of our experiences in such terms. It is hardly surprising that we do, if our principal interest in perception is in gaining information about the external world and not about the mental states involved. But I am arguing that those states do have recognisable qualities, and that therefore, in principle, terms can be invented for describing them. Rather than invent such terms, however, it is easier to use existing linguistic devices. If the argument of this thesis is correct, then this carries no more theoretical commitment than mere convenience.

Steve Voss suggested in conversation that a version of (3) has the advantage of economy over the explanatory postulate of the reality of the physical world.

The version is Descartes' Evil Genius hypothesis.

It has, I suppose, the advantage of economy of number of entities - one entity against many - but it surely suffers in comparison of predictive power and "coherence" of theory. We would have no idea what sensations the Evil Genius would give us next, or in given imagined situations. Nor would the sensations we have already had in any way hang together, being as they are each of them the result of an arbitrary and logically unconnected decisions on the part of the Genius.

We might try to postulate a whole character for the Genius to provide motives for its giving us just these ideas (but imagine what it would have to be like to invent any sort of detailed Grand Plan for what we see around us !), and then try to invent a "physics" for how the Genius goes about causing what he does. We might come up with something like the Dr. Doom story. An interesting point here is that the more causally structured the explanation, the closer it begins to look to an external physical world anyway.

The advantage of the physical objects hypothesis is that it allows us to do science, in that it is a necessary part of any science that we know today. If we grant RR, and so grant a certain picture of perception, then part of what science can do is to give an account of the causal mechanism of that perception.[26] In so doing, it will be able to explain the regularities in our experiences and moreover one day be able to explain all the departures from regularities as instances of other

---

[26] Even if it is not wholly successful at giving a totally physicalist account.

regularities - it can do that to some degree now. That gives it considerable coherence, demonstrating that the regularities in our experience stem from some not too large set of laws of nature together with facts about the set-up of human perception. Predictions, such as the prediction that when the sense data of the clock hands next stands at 6.30, the sky will be light, can certainly be given a reasonably detailed and believable explanation.

The hypothesis of physical objects ~~hypothesis~~ does not do all this by itself, but it is a necessary part of all the likely candidates (given that we accept RR) for what will. It seems to me that anyone who rejects the necessity to explain our experience rejects science and with it the physical objects hypothesis; and conversely to accept the desirability of explaining the order in our experience is to require that the explanatory hypothesis be potentially causally rich enough to do so, and the only ones we have that are remotely strong enough are ones all of which entail that there is a mind-independent physical world.

Notice the empiricism in the quote from Armstrong. We "degrade" bodies into hypotheses, which we can only "indirectly confirm". It is a mistake to think that because there is the smell of an hypothesis around, the knowledge in question is somehow second-class knowledge. There is surely no doubt that one can reach sure and certain knowledge by using hypothetical methods.[27]

---

[27] Maund has also interestingly challenged Armstrong's claim that it is improper to have a "lurking doubt".

Objection 3. On RR, we could not have the concept of a physical object; we could not "think" it, or "understand" it.[28]

If the last objection derived its force from empiricist epistemology, this one is ultra-positivist. The objection would seem to allow us to understand only that which we can be directly aware of. That restrictive picture of knowledge is surely mistaken. In this regard, it is interesting to look at the concepts clustering around "physical", such as "spatial", "temporal". If we say that our sensations have a fairly stable set of spatial causes, what must we mean by this? If we ask what physics means by objects being in space, we get quite a complicated answer. It is for objects to be somehow embedded in a manifold with very complex properties involving at least (multi-) dimensionality, and a continuous ordering within each dimension. (I do not want to beg the question of the absolute-relational nature of space, though I believe it to be absolute. I will talk absolutist talk and hope there is a suitably neutral way of expressing my remarks.) Now nobody supposes that we are somehow directly aware that space has these properties. The world is postulated by physicists to be like this. So if we want to say that there are spatial objects and we want a show at this claim's being true, so that we will want to mean by it what the physicists mean, we will have to admit that the claim has the status of an hypothesis. (This gives extra weight to the conclusion of the answer to the previous objection too.) This seems to me to be an adequate answer to the objection, unless the objector is prepared to claim that they cannot

---

[28] Martin Lean 1953 p.99.

understand the concept of spatiality that physicists are working with. To which the answer is, presumably, so much the worse for the objector's understanding mechanism.

Objection 4. We take it for granted that what we perceive are external objects, but the theory (RR) must say that this is an illusion, so the theory is counterintuitive. This objection has been made by, among others, R.J. Hirst.[29]

This calls for considerable comment. The first thing I want to say is that it is an easy trap to fall into, to think that mediated awareness is not real awareness; that real awareness can only be direct awareness of the external object or, in the case of RR, the mediating object or state. I simply deny that this is so: the process which, according to RR, is perceptual awareness of external objects seems to me to be a perfectly respectable sort of awareness. Substitute "knowledge" for "awareness" (as we have been allowing all along) and this point seems clear. Perhaps the objection derives some of its force from presupposing some sort of Acquaintance theory of awareness. Alternatively, perhaps it derives some of its force from some sense of "aware" which is not cashable in terms of "know".

Again, to repeat the point of the answers to the last two objections, it is a mistake to suppose that hypothetical knowledge is somehow second-class. One source of this confusion is traditional empiricist thinking: real knowledge must be certain, certain knowledge must

---

[29] Hirst 1959 pp.172-6.

be free from error, and conclusions established by inductive or hypothetical means are always open to error.[30] It is no part of the position of this book that any of the sources of our knowledge are incorrigible.

However, there is something else in the objection. Namely, that what we are aware of in perception carries with it the idea or suggestion of externality. It is difficult to believe that what we are aware of in perception is not wholly outer, or distinct from ourselves and our states. It is difficult to believe that anything gets in the way between the table I am looking at now and my awareness of it. Is it not the case that all of what I perceptually know now, or even when I switch my mental-state introspector on, is propositions about external states of affairs? Yet the theory being advocated would have us believe that we can come to be in the position that part of what I can know is propositions about the mental.

I am sure that this is one of the considerations at the bottom of the feeling of intuitive rightness that Directist theories have. In normal perception, everything does seem wholly outer, so the idea that part of what we are aware of is (or can be) our mental states, is counter-intuitive.

We should remember, however, that sometimes this intuition just <u>is</u> wrong. Sometimes full-blown

---

[30] And so since there is knowledge of contingencies, there must be incorrigible knowledge of sense data. Thus hypothetical knowledge of the external world cannot be first class. It is curious the extent to which some objections to Representative Realism derive from a philosophical position which was eager to embrace sense data.

hallucinations occur and we do not believe that we are hallucinating but really seeing. Out state gives all the impressions of externality. Yet in hallucination we can come to know that part of what we know things about is not something "out there". So our intuitive feelings can sometimes be wrong.[31]

So doubt such features of our perceptual-sensations as their quasi-spatiality and constancy contribute to this idea of externality. Perceptual-sensations display a space-like ordering, in that such states have their parts in an ordering which is irreflexive, asymmetric, transitive (and perhaps dense) and which displays some of the topological features of dimensionality. Our (postulated) physical space is much the same, so what a natural thing to think that what we are aware of is in physical space, especially when there does not seem to be any room for two different physical spaces in the universe ! Furthermore, sitting in a room looking at the furniture which is not moving gives an air of constancy or immovability, which seems foreign to such ephemeral entities as sensations, and more the hallmark of the external. We will return to this point about the topology of our mental states in the next chapter, as it provides a reason for denying materialism.

So it is natural to think of all of what we are aware of in perception as external. I am sure also that it is <u>prudent</u> to do so. The creature which evolved

---

[31] We might say that in hallucination we do always have beliefs about our states in that we believe that <u>this</u> thing is in space-time, where "this thing" denotes some mental item though we do not know that it is mental. I will not explore this possibility.

without a healthy paranoia of external dangers being revealed by its perceptual-sensations, would not be too successful. At a primitive level of existence, it is the beliefs about the external world which are more important for survival than beliefs about inner states. Even though the beliefs that we have about the external world can be examined (with an eye to justification), as we have done in this section, it is clearly an evolutionary successful trait that we should have a natural propensity to acquire such beliefs without intervening reflection and justification: so much so that it might be called a sine qua non of successful evolution. So the natural tendency to acquire such beliefs itself probably adds to the illusion of externality. But, as I have been arguing, there is a sense in which it is an illusion. Perhaps the point is best put this way: the illusion is that what we are unmediatedly aware of in perception is external. I am not claiming that it is an illusion that some of, and sometimes all of, what we are aware of, mediated or unmediated, in perception, is external.

This complete the objections to RR that I wish to discuss here. It remains to sum up the discussion of perception.

10. Conclusion.

Why should we believe the theory I have advocated here? We might try advancing the whole theory as an hypothesis. The trouble with this approach in our case, is that the theory does appear to have somewhat unpleasant consequences, not the least of which, I will admit, is a taint of dualism. So as an hypothesis it certainly falls foul of Ockham's Razor. Thus without

some more direct arguments to back it up, we should go for theories better suited to materialism. We might try criticising all alternative theories. But a demonstration that all alternative theories are false will amount to a proof that our theory is true. But then alternative theories are often only shown to be counterintuitive rather than false, so our method is often a mixed one. In any case, the middle ground between treating a theory as an hypothesis, and proving it true by some valid argument with known true premisses, is hazy. I take it that my reasons for believing the theory are not on the hypothesis end of the continuum.

Some of the time in perception we know that we have perceptual-sensations (perhaps not under that name). Certainly, but a lot needs to be said. The sort of cases central to the argument are hallucinations. The real job consists in establishing that what we (can) know about in such situations, is not merely a belief or tendency to believe, or arrested belief; that what we are aware of is not something belief-like. I believe that this has to be done otherwise there is no real case against DR and the simplicity of the latter carries the day. I have tried to do this job.

Hallucinations and afterimages are not perception, and so the point of the previous paragraph does not show that sometimes <u>in perception</u> we (can) know that we have perceptual-sensations not belief-like. But all of the time in perception part of what is occurring is also what occurs in hallucination and afterimages even when we are not fooled to the point of having no tendencies to have beliefs about the external world. Perhaps in the end I just have to appeal to you to see that what I say is

true. Close your eyes, open them, then close them. Something comes into existence in the period when your eyes are open, which goes out of existence when you close them. Say "occurs" if you do not like "exists". I claim that what comes into and goes out of existence, is the sort of thing which goes on all the time in visual perception, and which we can come to know about when our attention is drawn to it. I further claim that this sort of thing, so closely involved in perception, is just the same sort of thing as what we are aware of in certain hallucinations: imagine looking at a blue hill, then unbeknownst to us the hill being destroyed and at the same instant there come into being an hallucination of an ~~identical~~ exactly similar blue hill, so that it seems to us that there has been no change. Clearly what we are sometimes aware of having in hallucination is what also goes on in normal perception.

"Sort of thing" is vague and messy, but metaphysics is sometimes vague and messy. However, we have previously been more precise, in that we have argued that our state has various properties, present in all of normal perception, hallucination and afterimaging, and which are necessary for perception as we know it. We have seen that these properties are not to be identified with belief-like properties, and that the states which have them are not to be identified with states which are just belief states. Give such properties and states names: "being a sensation-of-red", "perceptual-sensations", etc. Then perceptual sensations are closely involved in normal perception. It remains to work out just how, and it has been part of the job of this chapter to do so.

Distinguishing between two kinds of awareness, awareness of outer objects in perception involving mediating sensations, and introspective awareness which, in central cases, does not seem to need mediation, commits one to the claim that not all cases of awareness are ~~the way~~ as I have been claiming visual awareness ~~is~~ to be. Introspective awareness is like the perceptual awareness of the Direct Realists. So why could not normal perception be like that? Well, it is not, but it could have been. So my theory is contingently true if true at all. Now whenever somebody makes such a claim, they are open to the question: what would it be like for your theory to be false? It is useful to try to work out answers to this question because you sometimes find that you are placing unreasonable demands, in trying to extract your conclusion from it, on the data you are using as premisses.

I do not think I could easily describe how experience would seem to the Direct Realist's Perceiver. Perhaps Keith Campbell's Imitation Man is the closest I can come: a man who directly and in a well-behaved causal fashion acquires beliefs about his environment. They just pop into his head, almost clairvoyantly.[32]

As an aside, people trying to understand what clairvoyance and telepathy might be like sometimes unreasonably insist that it be like a sixth sense and involve a new sort of sensation, or that it be a merging of minds with trouble about personal identity. I do not see why a telepath could not simply have beliefs pop into his or her head, and the beliefs be true, and there be

---

[32] Campbell 1970 Ch. 5.

some unknown and perhaps nonphysical causal mechanism to make the true beliefs into knowledge.

Return to the Imitation Man, who is not telepathic because the causal mechanisms involved are not too far out. Perhaps things would seem to him the way the world would seem to a blind clairvoyant who just knew directly the colours and shapes and positions of objects.

In this connection, we must beware of arguing for Direct Realism on the grounds that our technology is not far from building a metal machine man to discriminate colours and shapes in its environment, and we could easily suppose that such a thing could be built without anything like sense data, and made to have illusions etc. I agree that we could conceivably make this robot man and that it conceivably could be made into a walking instance of Direct Realism. But I do not agree that this would show that we are similarly constructed. In particular, I think that the nearest the metal machine man could come to hallucinations and afterimages would be to have a series of suppressed tendencies to believe that the environment is a certain way. I do not agree that this is how we have hallucinations and afterimages.

This completes our discussion of the philosophy of perception. I now want to return to the main theme of the later part of this book, which is: what to identify being a sensation-of-red with? We have argued that it is not to be identified with any causal or relational property of sensations-of-red, nor with any belief or suppressed tendencies to believe. There is, it seems to me, just one avenue left to try to save physicalism. This is, to identify the property with some physical property the nature of which we do not today know, but which

science can be expected ultimately to reveal. We will turn to this in the next chapter and argue that the identification cannot be made.

CHAPTER ELEVEN. ARE PERCEPTUAL-SENSATIONS PHYSICAL?

1. The Argument For Physicalism.

We must ask what sort of property being a sensation-of-red is. We have argued that it is neither a tendency nor a belief-like property of our states, and we have argued that it is not a causal property of them either. What is left?

Nothing is left that I can think of to save physicalism unless it is that being a sensation-of-red is just some as-yet-unknown property of brains. Now it is notable that we cannot tell just by paying attention to our sensations-of-red that the relevant property is something neural. But, as we have said before, that should not by itself prevent us from making the identification.

We have already argued that there can be contingent identification of properties. It is not clear that what we have here is a (purported) instance of <u>contingent</u> identification or not. We invented a name for the property in question, and that might plausibly be thought enough to show that the identity statement will be contingent. On the other hand it might be argued that paying attention to the property would reveal enough of its essential nature to show that it must be identical with a physical property. Whether this would be an instance of non-contingent identification of being a sensation-of-red <u>under some description</u> is arguable. It might be argued, for instance, that the conditions for fixing the name "being a sensation-of-red" were such as to tie <u>logically</u>

certain aspects of the property to the name. This would give a rationale to speaking of those aspects as (part of) the essential nature of the property.

This question is independent of the question of whether the property already has a name in English. We have not investigated the question of which property "x is red" is a predicate for, and "redness" is a name for. Certainly they have a sense in which they stand for an ordinary physical property of physical objects. In my view they are not univocal in this, sometimes also being used when it is the property being a sensation-of-red that is up. The sort of distinctions which are necessary before a person can come to know just which properties they or others are denoting on a given occasion, sometimes presuppose a level of theoretical complexity that most people do not have. A person, too, might be quite confused when you start asking them about their "sensations of red" (no hyphens), but after a while catch on and use this English phrase without any apparent shift in its meaning.

So the question of how to analyse "being a sensation-of-red" is problematic. One thing we can say, though, is that even if an analysis were favourable to physicalism, this would not demonstrate physicalism. To repeat a point made previously, it might be that our language is not adequate to the facts, and that when the facts are grasped we will have to invent linguistic forms for expressing them.

How should we decide this question, if we were in possession of the hypothetical analysis? Look at the property itself and see if its being instantiated is incompatible with physicalism. See what we can adduce

in the way of facts about entities that have the property. For instance, see if there is any barrier to identifying the property with some neural property.

This last point enables us to cut across the question: contingent or noncontingent identification? For if we can successfully argue that being a sensation of red has a certain property, say a certain causal property, and also argue that just one property, a physical one, has this property, then we can make the sought-after identification without being involved in questions of contingency, or questions of analysis.

This presents a last, and very formidable, argument in favour of physicalism. It seems undeniable that the possession of being a sensation-of-red is causally relevant in various ways in various situations. It is equally undeniable that Ockhamist-type unity of science reasons should lead us to expect reasonably that a physical theory will come to explain successfully all human behaviour. That is, we should expect that there is just one property the possession of which is causally relevant in just those ways in just those situations, and that it is physical.

Is there anything at all that a dualist can say against this? If a dualist cannot point to our lack of knowledge of the neural properties of our brain to back him or her up, and if a dualist cannot point to any impossibility of analysing the name standing for the property being a sensation-of-red to back him or her up, then there would seem to be only one possible course: to try to show that, from what we can find out about the property, its nature is such that it is not identical with a physical property.

It is not absolutely clear to me how to describe the method of argument that I have in mind, and so I am not going to try, but rather plunge straight into it.

2. Properties of Our States And Properties of Ourselves.

We have characterised being a sensation-of-red as a property of our states, but let us be clear that our states might have such a property just in virtue of something not identical with any of our states having another property. In fact, this follows from a very reasonable interpretation of what it is for one of my states to have a property. We introduced "state" talk by noting that our state on certain occasions was like in some respects and unlike in other respects, our state on other occasions, and we said that the interpretation of likeness in a respect, was the possession of a common property. So we were committed to there being properties of our states (and why not, if we quantify over states?). But we did not say what sort of things properties of our states or, indeed, properties in general are. The idea was introduced and left deliberately neutral, our only commitment being to the fact that they existed, were some sort of universal, and were whatever it was that made possible multiple predication. Now the most plausible interpretation of what a <u>state</u> is, it seems to me, is that it is some unity of properties. (The word "unity" is used here to be neutral between words like "set", "whole", "collection".) Not any unity of any properties of ourselves will do, it might be said, for if we do not restrict which properties go to make up my state, we will have the (Hegelian?) consequence that a specification of my state is a description of the universe. We will not investigate this matter, but just assume that

the problem of which restriction to place on such properties has been solved. Nevertheless, a specification of my state at a given time is completely given by a (suitable) list of the properties that I have at the time, and so we can say that my state is the unity of those properties.

This gives us a way in which a complex state of a person can share a property and fail to share another property with another complex state of the person. We can say that state s_1, of person p_1, has properties F_1 and F_2, and s_2 of p_2 has F_1 but not F_2, in virtue of s_1 being a unity of properties of p_1 which includes ϕ_1 and ϕ_2, and s_2 being a unity of properties of p_2 which includes ϕ_1 but not ϕ_2, where ϕ_1 and ϕ_2 are properties of persons rather than their states. For example, we might say that yours and my states have the property, being a having of a red afterimage, in virtue of you and I having the property, having a red afterimage.

In a similar fashion, it might be that a person possesses a property in virtue of some entity not identical with the person possessing another property. The other entity might, for instance, be a part of the person. We can say that a person possesses the property, having a red arm, in virtue of a part of the person, the arm, possessing the property redness.

This is what I want to suggest is the case with the property "being a sensation-of-red". As we introduced it, it was a property of our states, which when they possessed it were said to be sensations-of-red. Now obviously we could say that our states were sensations-of-red in virtue of ourselves possessing the

property, being a haver of a sensation-of-red. But that is not quite what I mean. There is something to be said about the structure of our sensations of red, which leads us to another property and another bearer of it.

I have all along been deliberately neutral about whether we afterimage or hallucinate. That was the reason for occasionally inserting "allegedly" in front of the relevant predicates. (We did not do this with "see", for it is too much to deny that we see.) Because I wanted to remain neutral about whether we actually afterimage, in order not to beg the question against Rorty, it was necessary to argue about the nature of the states that we are in on those particular occasions, (which undoubtedly occur) which we identify by saying that they are the occasions about which there is dispute over whether we afterimage. I have argued to the point where it has been shown that those states have certain properties, which can be known to us, and which are, perhaps, physical. Just now I offered an account of what it is to be in a state, and what it is for the state to have certain properties. The argument to be given will not rely on assuming this account: it was included to illustrate a possible way in which a state of a person might be said to have a property in virtue of the person's having another property. The point I want to make, though, is that there is no particular problem about there being properties of our states. Since they are our states, we could just as well have centred our discussion on certain of our properties. The real question is not about whether it is our states as opposed to our properties that are material or not. It is rather about the nature (physical or not) of what it is about us on the disputed

occasions. So far, the argument has established that we can come to have certain knowledge about ourselves on such occasions, and that what we know about ourselves is that we or our states have certain properties, and that these properties are neither causal nor belief-like properties. The effect of Rorty's arguments has been the inconvenience of being unable to conduct the discussion as a discussion of the nature of our afterimaging or hallucinating.

3. The Topology of Our States.

It is clear that our states on the relevant occasions are very complex. There are many things about us, happening to us, going on in us, on those occasions, heartbeat, food metabolism, breathing, neural activity. I want to concentrate on just one aspect of our states or ourselves on such occasions. I expressed that aspect before by speaking about certain properties of our states: being a sensation-of-red, being a sensation-of-a-square.

Now what it is that we (can) know about ourselves on such occasions is a series of facts with (at least) one very interesting structure. It might be called a "quasi-topological" structure. It is a structure, i.e. a set of relations, on the facts or propositions we know about ourselves. We might also have said that it was a structure on a certain class of the properties which we possess on those occasions. It is better to talk in this way because it allows for the possibility of discussing our possession of those properties even when we do not know that we possess them. The way we established that we possess these properties was by an argument using the facts we know

or can come to know about ourselves on those occasions, but we have seen that the possession of the properties is not necessarily conscious.

We first notice that the structure contains a systematic class of non-identities between our properties (or states). The state when we (allegedly) hallucinate just one figure, a square (suppose that all the figures we are talking about are red), is not identical with the state when we hallucinate just one figure, a circle. The property we possess when we hallucinate just one n-sided polyhedron is not equal to the property we possess when we hallucinate just one m-sided polyhedron, when $n \neq m$.

If we are to identify each of these properties with some internal physical representation, then the physical representations (properties) will need to have a corresponding system of inequalities. If this were all there was to it, then perhaps the identification could be made. Mere equality and inequality, however, hardly begin to do justice to the relations between the properties we are discussing.

For example, we can hallucinate at t_o a triangle, and then at t_1 hallucinate a rotation of that triangle (that is, a rotated triangle. We can, of course, also hallucinate dynamically a rotating triangle. That is not what I mean). Now to represent the relationship between these two properties physically, we should need some aspect of the relation between the two which somehow represented the rotation. I do not mean that the property associated with hallucinating a triangle at t_o is a rotation of the property associated with hallucinating a triangle at t_1. That does not even make sense. I mean

that there is a systematic relation between the properties which is marked by the rotation of the internal object of the hallucination - it is one discernable dimension among others in which the properties can vary - and which therefore needs marking by some systematically variable feature of the physical properties with which to identify them.

Compare this way in which our properties vary with another way: translation. We can hallucinate the translation (movement from one part of the visual field to another) of a triangle. In both this case and the previous one, something stays the same (that it is the state associated with the alleged hallucination of a triangle), and something varies, but it is a different respect which varies when there is rotation from the respect which varies when there is translation. And it is something which systematically varies (for there is a systematic series of rotations), so we will need for the identification a series of neural properties, and to keep it from being just an arbitrarily selected series of physical properties, there must be some definite non-arbitrary relation between them. I take it that just to include the neural properties in some set of n-tuples would be an arbitrary way of making the relation, for the properties can be included in many sets of n-tuples. So we will need some definite feature of the properties - some definite feature of ourselves - which systematically varies during rotation, and some definite other feature systematically varying during translation. Thus, as I said before, we need some aspect of the relation between the two properties which represents the rotation. So we cannot rest content merely with identifying different properties with

different physical properties. The physical properties need to have a definite structure of their own, corresponding to the ways in which our allegedly-hallucinating properties can vary. At the risk of labouring the point, simply to think of hallucinating a triangle, then hallucinating the same triangle rotated through 45^{o}, then hallucinating the same triangle rotated through another 45^{o}, as being in three distinct states (as possessing three non-identical properties) fails to mark the partial similarities between the states and the systematic variation between them. These partial similarities and various systematic variations must be marked by corresponding similarities and systematic variability in their physical "analogues".

Thus an hallucinating of a triangle at a certain point in the visual field and with a certain orientation is not a simple thing, but contains information about the nature of the figure, its position and its orientation. Any one of these can vary while the others stay constant, so whatever physical state or property represents this fact about ourselves must map at least these features.

In belief-like items there is a ready-made map (provided we could map beliefs into physical structures). To every distinct feature of an hallucination we could associate a part of the proposition describing the contents of the hallucination (i.e. a proposition not about hallucination but about geometrical shapes) and then simply place an "x believes that" operator in front of the proposition. But we have argued that the tendency to have such beliefs need not always be present in these circumstances.

To carry on with our brief outline of the complex quasi-topological structure of our hallucination-properties, there can be such relations as congruence, and similarity between the figures in the hallucination, and congruence of some parts of the figures but not of others. These aspects of our hallucinations should need marking by relations between their physical bases.

In addition to these relations between alleged-hallucination-states, there are relations between figures within the one hallucination state serving to distinguish hallucinatings. We might term these relations the "internal topology" of an hallucinating. For instance, something distinguishes allegedly-hallucinating a square to the left of a circle, from a square above a circle, below a circle, and to the right of a circle. Something distinguishes hallucinating two squares from hallucinating three. In fact, more or less[1] any configuration of shapes which can be achieved in a bounded connected Euclidean region of (at least) two dimensions, can serve to distinguish hallucinating-properties and afterimaging-properties. Identifying such properties with a system of physical properties, therefore, should require a system of properties of complexity great enough to match adequately these various relations between hallucinations.

There does not seem to me to be any a priori reason why a sufficiently complex series of elements and properties adequate for encoding these differences

---

[1] Leaving aside the question of whether we could be said to hallucinate e.g. a 10^{50} - sided figure - i.e. the question of whether the internal topology of hallucinations really is that of a bounded, continuous, two dimensional, connected, etc. region.

should not exist in the brain. I say this with one exception in mind: continuity and density (Sellars raises this problem[2]). If the "internal topology" of our hallucinatings really is continuous, so that there are various continuous series of hallucinatings to encode (and hence c of them), then the brain which is discrete at the neuronal level (even at the quantum level !) will be unable to achieve this. Similarity with density. If an hallucinated line of finite length really is made up of a dense series of elements, then to represent all the possible different length lines would need an infinite sequence of different properties, which is impossible for a discrete brain. I leave this problem with the observation that a discrete serites can seem to be dense if the difference between members of the series are too small for us to be aware of them, for instance in movie projection.

Leaving this aside, if all that is needed is to represent a finite amount of information, then there is surely no reason why the brain should not achieve this provided the amount of information is not too great. I do not think that I have an argument for the conclusion that we have too many different hallucinating states, or some states are too complex, for the brain to represent.

4. The Relational Nature of Visual Hallucinations and Afterimaging.

When we (allegedly) hallucinate two triangles side by side, in normal conditions there often seems to be two (triangular shaped) objects in front of our eyes. We might also have said that there appears to be two objects

---

[2] Sellars 1963 p.191; Aune 1967 Ch.9.

in front of our eyes. To repeat something said earlier, at the very least this involves under normal conditions the belief or tendency to believe that there are two objects in front of the eyes.

Now remove the tendency to believe that there are two triangular shaped objects out there in the world. Suppose that there is a case where we have no such tendency: we know there is nothing there and we calmly introspect our alleged-hallucination-state. Let us say that the triangles are coloured red and that there is nothing else in the visual field, it being all black. Do you not note that there still seems to be two (triangular shaped) things in existence? I have not said that there seems to be two objects in front of the eyes, out in the world as it were, for that would commit us to the tendency to believe that the triangles are in the physical world. It seems clear to me, however, that there does in such a case seem to be two (not three or one) objects, or regions, triangular shaped, in existence. Suppose one of them winks out. How it seems, is that one of them has gone out of existence. Suppose three more now appear side by side. It seems that three triangular things have come into existence.

The "it seems" here can only, it seems to me, plausibly be rendered in terms of beliefs and tendencies to believe. When it seems that there are two triangular regions in existence, we believe or tend to believe that there are two triangular regions in existence. And when you inspect the contents of your hallucination, do you not come to have precisely this belief or tendency to believe?

I have not said that the belief that there are two triangular regions in existence will invariably accompany being in the alleged-hallucinating-of-two-triangles-state. It has already been argued that in principle we can afterimage or hallucinate an A, and have no beliefs either that an A exists, or even, for reasons of corrigibility, that we are afterimaging or hallucinating an A. In this case, I am not arguing that the belief must be present. I am inviting the reader to examine the nature of his or her hallucinating or afterimaging[3], and note what they come up with.

So if you examine yourself even when you lack all tendency to believe that there are two physical triangles in the world, you will have a tendency to believe, perhaps a full belief, that there are a pair of triangular objects or regions.

Now I will say that this belief is in fact knowledge. More exactly, I will claim that the belief is true, and leave unargued for the contention that if it is true, it will count as knowledge in virtue of there being a suitable causal relation between the belief and what makes it true. I will take it that this latter condition is satisfied.

Why should we say that the belief is true? A tough-minded answer is that if you attend to the contents of your mind when you are allegedly-hallucinating two

---

[3] If the reader has trouble imaging this case, one way of placing yourself in the right sort of state is to draw two bright green triangles on a piece of white paper, stare at them for some time under bright light, then go into a dark room or close your eyes.

triangles you will see that it is manifest that this is what it is like. Two triangular-shaped regions really do exist. Look and see.[4]

I think that this is correct, and sufficient to establish my conclusion. I realise, however, that many philosophers would be unwilling to accept it. I will therefore give further arguments for the same conclusion.

I want to say that if the belief that there are two triangles in existence is not knowledge, then neither are many of the other beliefs that we have already held to be knowledge. We have agreed that when we were in the allegedly-hallucinating-a-red-object state, we could tell that we were in a state like in some respect other allegedly-hallucinating-a-red-object states, allegedly-having-a-red-afterimage states, and seeing-a-red-object states. We remarked later that the same points about what we could tell concerning similarities, differences, and respects, could be made about the states when we allegedly hallucinate a square, allegedly have a square afterimage, and see a square. We used these facts about what we could come to know as the basis of our argument for the conclusion that the relevant aspects of our states were not to be identified with belief-like items. We can of course make the same points about (allegedly) hallucinating a pair of triangles, etc. We can tell that these states are similar. That is, if you attend closely to your state, you will come to know that it is similar to various states. But this is precisely the method I am recommending that one uses in coming to believe that there are in existence a pair of triangles: careful attentive

---

[4] Metaphorically speaking, of course. c.f. Philosophical Investigations, section 66.

introspection. I do not claim that careful attention will always give accurate results. I do claim that if we hold that it can accurately tell us very complex information about similarities, differences, similarities of respect, differences of respect and so on, and that it does not accurately tell us that there are a pair of triangles present, then we should need to know a relevant difference in the two cases. The difference should be relevant to showing why we can count the one complex set of beliefs as knowledge, but cannot count the other belief as knowledge. I can think of no difference which even begins to be relevant to showing this.

This, then, is my second argument: if we fail to count the belief that there exist two triangular regions as knowledge, then we ought also fail to count the various beliefs that we are often in states similar to one another as knowledge. This latter seems too much to hold as has been said earlier (Chapter Nine).

Perhaps one source of reticence to accept the conclusion of this argument is the feeling that if it were true, the things in question that are triangular would be funny entities. I do not mean that they would be funny insofar as they would be nonphysical, but, rather, funny in that they would be only part of the story concerning our state. They would not be in any way substantial items embedded in a manifold, but rather more like parts of a manifold, the visual field. There is something not quite right about talking about parts of a piece of space as existing or for that matter of talking about the "visual field" as a piece of space and hence existing.

There are two answers to this. The first is that I am not disturbed by the possibility that space is

real. I think that it is real, and that it has real parts. The second is that our argument does not establish that the triangles, or any visual-field-type entity of which the triangles are parts, are in any way insubstantial in the way that space is. In fact it does not even establish that the triangles or the visual field are not physical objects. Even if we had established this much, it would still not follow that the triangles and the visual field were not constructed out of spiritual stuff with a (nearly) spatial topology. There would presumably be no contradiction in talking about such stuff and supposing it to have parts made up of the same stuff and triangular shaped.

Might we not try to analyse away, or adverbialise, that which it is being claimed we know when we attend to our hallucinatings? The only function that such steps could serve here would be to deny what I am claiming we know, and hence to deny what is the case. If my argument is correct, then such moves must be mistaken. It is worth seeing in this perspective the analysis and adverbialisation methodologies which we discussed earlier. They can sometimes serve the function of dealing with troublesome predicates. But on the other hand, they must remain within the limits laid down by the facts that introspection reveals.

5. Explaining Entailments Between Afterimage Predicates.

If under the stipulated conditions we do know that there is a pair of triangles in existence, then much that is otherwise puzzling about the phenomena of hallucinating and afterimaging immediately becomes unpuzzled. Note first that the argument immediately

generalises to the hallucination or afterimaging of any geometrical figure. Then, we have a ready-made systematic (indeed, mathematical) account of the various complicated orderings of properties and states of ourselves which we have described. They have these orderings in virtue of the properties being relational properties between persons and certain regions with the various topological orderings possessed by sub-regions of a certain sort of topological space. To take a particular case, we have an immediate explanation of a point raised earlier in connection with Jackson's work on adverbs: that an hallucination or afterimage of a single figure, a square, cannot be an hallucination or afterimage of a single figure which is a circle. Indeed, it is a superior explanation to one which, like Jackson's, proceeds by way of arguing for the reality of afterimages. For that argument must also give enough of the semantics of "square" and "circle" considered as predicates of afterimages, to guarantee that the one includes the other. Here we have a cut and dried reason for that exclusion; "square", "triangle" are intended to have their usual geometrical meanings, and squares cannot be circles.

Indeed, we have an explanation of all the "geometrical" entailments between hallucination and afterimage predicates. For example, if we hold that to hallucinate a square is to be in a certain sort of relationship to an actual square thing, then we have a framework for the semantics of hallucination predicates. If we make the further assumption that the relationship in question is extensional and so allows substitutivity of identity, at least with respect to a suitable class of terms, then we can immediately deduce that to hallucinate or

afterimage a square is (at least) to hallucinate or afterimage a quadrilateral.

It might be objected that it is not obviously an advantage to be able to explain these entailments, because it has not yet been established that hallucinatings take place. The argument has been conducted without assuming that afterimage and hallucination predicates are instantiated, it being maintained only that on the disputed occasions people are in special conditions that they are in a position to know something about. In reply to this objection, let me say that we have established enough about those conditions and what we know about them to make one wonder what advantage would be gained by denying the instantiation of the predicates. This is especially pertinent if we remind ourselves that inability to analyse away the predicates does not prevent the possibility of their teeth being drawn by a simple identification of the properties they stand for with physical properties. It is certainly the case, furthermore, that afterimage and hallucination predicates give us a convenient syntactic mark of the presence of complexes of features of ourselves. The description of the internal object of the hallucination contains various terms to which, even if we do not agree there exist things bearing properties denoted by those terms taken in a literal sense, there correspond discernable and discernably distinct aspects, respects, properties, features, what have you, of ourselves and/or our states.

The usefulness of this terminology leads me to make the assumption that afterimage and hallucination predicates are instantiated, and stand for just those complexes of properties of ourselves that we have been

at pains to elucidate. We will keep a watchful eye on the possibility with this assumption that too much metaphysics can be extracted from it. Insofar as you agree with what is assumed, i.e. that humans do afterimage and hallucinate, you will presumably agree with the point that we have here an undoubtedly ready-made explanation of the puzzling entailments.

6. Is It a Better Explanation Than Its Rivals?

That is a point about how it explains certain entailments. It is not a point about its being a <u>better</u> explanation of those entailments than certain other possible explanations. In order to make <u>that</u> sort of point stick, we should have to spell out in considerable detail the structure of the explanation and its rivals, and that is not what I wish to do here. However, I am inclined to believe that it is a better explanation than at least three of its rivals. In this section, I will sketch extremely briefly why I believe that it is a better explanation.

(1) Take the first rival I have in mind: hallucinating or afterimaging a figure with geometry G is having (possibly repressed) tendency to believe, or a full belief, that there exists an object with geometry G. It is important to distinguish this account of hallucination and afterimaging from the Direct Realist account of visual aberration, according to which such aberrations involve the belief that there is an object with geometry G <u>in the world</u>, in front of the eyes perhaps. Now, if I am right, this latter account can be defeated by the arguments of the previous two chapters. But obviously we cannot cite the manifest absence of any tendency to

be deceived about the existence of objects in front of the eyes as an argument against the former account, for different beliefs are up for consideration.

Nevertheless, the new account has two difficulties with it. One is that it is implausible in its own right. The other is that it is inferior as an explanation of the entailments to the view I am defending.

The first difficulty is this: as I said before, I do not claim even that the tendency to believe that there are two triangles in existence invariably accompanies hallucinating or afterimaging two triangles. The corrigibilist position defended in Chapter Five allows the possibility that we hallucinate or afterimage, and not be aware of it. Furthermore, this seems to be quite a frequent occurrence at least with afterimaging. The phenomena of suddenly noticing one's afterimages does occur. This gives a reason for thinking that an afterimage of an X is not invariably accompanied by the tendency to believe that there is an X in existence. Furthermore, the Argument from the Speckled Hen lends weight to this point. It was argued in Chapter Ten that inspection of our mental states can take time, and that before inspection it is implausible to think that we necessarily must have beliefs corresponding to all the introspectable features of our mental states. But that is an argument which goes through independent of whether the beliefs are beliefs about there being certain objects with certain properties in the physical world, or simply beliefs about there being certain objects with certain properties in existence.

The second difficulty is that beliefs do not have a strong enough semantics to give us the required

entailments. There is always the possibility, even with the most obvious of entailments, P entails Q, that a sufficiently irrational person might believe that P and fail to have a tendency to believe that Q. For instance, it is surely possible for a person to believe that there exists a Star of David (which, perhaps, they are hallucinating), but fail to believe that it is a six-pointed star. On the other hand, surely if one hallucinates a Star of David, then one hallucinates a six-pointed star. This argument does not establish that hallucinations are not identical with belief-like items. There is no reason why substituting for identicals need preserve entailments. It does establish that the belief-account of hallucinating does not give an explantion of the entailments.[5]

(2) The second rival for the explanation of the entailments is some sort of Rennie-style semantics for afterimage- and hallucination-predicates designed to make the entailments come out, coupled with the claim that the various inclusion, exclusion relationships in the semantics are dictated by the meanings of the predicates in question. This has a ring of ad-hocness and lack of system about it, although I am very unsure about this point. For instance, there is no a priori reason why "square" in afterimage-contexts and hallucination-contexts should be associated with just one semantic unit, and thus any claim that it is would seem to need <u>further</u> explanation. A Rennie-style semantics which merely gave inclusion, etc., relations between predicates, even if it did associate "square" with a single semantic unit, appears not to give this further

---

[5] It would also, therefore, seem to establish that the belief-account does not give us the meaning of the hallucination-predicates.

explanation. Such a Rennie-style account seems to be, though I am not sure how to show it, just the claim that a semantics strong enough to guarantee the entailment will guarantee it.

(3) The third rival I have in mind is Jackson's argument for the existence of afterimages (Chapter Four). Now the claim that afterimages are real does not, as I said before, by itself guarantee the entailments. It must be supplemented with enough of the semantics of "square", "quadrilateral" considered as predicates of afterimages to make the entailments work. As a particular case of this, it might attempt to argue that "square" is to be taken literally in such contexts. But it is clear that any such argument must be additional to the claim that afterimages are real. Furthermore, any argument for that conclusion is an argument for the conclusion being urged here.

7. The Colours of Our Afterimages and Hallucinations.

I will not pursue this interesting side-line further. It is important to see that what has been offered here is by no means a sketch of the full semantics of afterimage and hallucination predicates. I have said that when we hallucinate a triangle we are in a certain relation to a triangular thing; something which is really a triangle. But we should not think that this holds for whatever terms come after the main verb. To hallucinate an A does not always entail that there exists an A, even if the argument given so far is correct. To hallucinate a cat does not entail that there is a cat (in my mind or anywhere else).

More important for my purposes, is the fact that it is by no means obvious that if we hallucinate or afterimage something red, then something red exists. We have argued that under these circumstances, we have a sensation-of-red, with the property of being-a-sensation-of-red, but this is a far cry from there being something red in existence. So, whatever semantics we give for "red" occurring in afterimage predicates, it cannot be precisely parallel to that for "square". We should like to have such a semantics, however, to extend the account of the explanatory virtues of the picture of afterimaging and hallucination.

I turn back to my stock weapon, careful introspection, and make some further wildly unsubstantiated assertions based on it.

If we hallucinate a red square, I claim that we are in a certain relation to something square, and that we can know about the existence of the square thing. The difference between hallucinating a red square and hallucinating a green square, has to do with a certain property being possessed by the square thing. If it is a red square, then it is one property. If it is a green square, then it is another. For it to be a red square rather than a half-red square, is for that property to be possessed by all the parts (or at least all the discernable parts) of the square thing. If it is half red and half green, then the square thing has two parts such that one of them has all its sub-parts possessing the one property, and the other has all its sub-parts possessing the other property.

These properties, which can be "extended" over the square, must not be confused with red and green.

(Nevertheless, I claim that we can know of their existence if we attend carefully.) I want to say, though, that they are importantly related to the predicates "x is a sensation-of-red (green)" and "x = being-a-sensation-of-red (green)". They are importantly related by virtue of the fact that when we have a sensation-of-red, what distinguishes it from having a sensation-of- green. is that some region specially related to us has that property which the square thing has when we afterimage a red square. That property is a property of some extended thing specially related to us just whenever we have a sensation-of-red.[6]

The property deserves a name, and it is a property closely connected to seeing a red object in the following way: when we see a red object, we have a sensation-of-red, and when we have a sensation-of-red, we are in a certain relation to a region with the property. Thus being in the relation to the region with the property is a necessary condition of seeing something red. It is not a sufficient condition because we can be in that relation to that sort of thing and not be seeing red, but afterimaging red or hallucinating red. The close connection prompts names: red*, green*, and so on. It should not be thought, of course, that because neologisms were necessary, the properties were ones with which we were hitherto unfamiliar.

---

[6] It may not be obvious what the shape of the extended thing is. This does not necessarily imply that the thing does not have a determinate shape: we might not be able to tell just by concentrating what that shape is. On the other hand, a patch of the property with streaky edges might not have a determinate shape for the reason that there is a region of indeterminateness about whether the property is possessed or not. I suggest this only as a possibility.

We are now in a position to extend the account of the explanatory virtues of the claim that in hallucinating a triangle we really are in a certain relation to a triangular thing. We add to the claim the further (above) story about red*, green*, etc. We ask what distinguishes seeing red, square, etc. from hallucinating red, square, etc. and afterimaging red, square, etc. The answer is basically a causal one (we omit conditions associated with the presence of beliefs, since we are not here interested in the entailments of seeing predicates, and we have argued that neither beliefs about the external world nor beliefs about ourselves need be present when hallucinating or afterimaging). To afterimage a red square is to be in a certain relation to an object which is red* and square, and for this state of affairs to have a certain sort of cause. If these conditions obtain, then, clearly, one is in a certain relation to a thing which is red* (and for that matter to be in that relation to a thing which is square). These latter states of affairs ex hypothesi have the requisite sorts of causes. Therefore, if one has a red square afterimage, then one has a red afterimage, and one has a square afterimage. The entailment of Chapter Four is explained. Obviously a similar argument will show that under these conditions one has a square red afterimage. Similar arguments will establish similar conclusions for hallucinating, if we distinguish that either by the operation of causes of certain sorts or by the absence of the operation of certain sorts of causes.

We also have an answer to Jackson's problem about how one could distinguish afterimaging two figures, a red square and a green circle, from afterimaging a red

circle and a green square. We have already seen that we can give an account of what it is to afterimage two figures, in terms of their being embedded in a thing with certain topological properties and related in a certain way to us. We need suppose further only that the square thing is red* all over, and the circular thing is green* all over. Since this situation amounts, with appropriate attendant causal relations, to having a red square and a green circular afterimage, it is evidently distinct from the situation where the square is green* all over, and the circular thing is red* all over, which is the condition of afterimaging a green square and a red circle. Indeed, I can think of no other satisfactory way of making this distinction.

The foregoing has been no more than a sketch, even of those parts of the semantics of the predicates that we did deal with. It omits a great deal[7], but it includes enough for us to see the great unifying explanatory power of our thesis. It explains much, and it explains it under the rubric of a single explanation. I take it that the explanatory virtues of the thesis constitute a third argument, and a strong one, for its truth.

8. The Physical Correlates of Visual Sensations.

So we have arrived at this point: to hallucinate and to afterimage geometrical figures is to be in a certain

---

[7] One interesting line to be pursued is the exclusion relations between the colours of afterimages. e.g. afterimaging just one red-all-over square entails not afterimaging a square which is any part green. One can see how the account might proceed (in terms of exclusion relations between red* and green*) but we will not pursue this matter.

relation to something with a certain geometry. Now we can return to our main theme and ask: can we identify these properties of ourselves with physical properties?[8]

Two answers to this question can be dismissed immediately: that the geometrical figure in question is a platonic object, and that the geometrical figure in question is a piece of physical space with the required geometry. Neither of these answers leads to a plausible account of introspection. In the first case we should have to say that when we hallucinate a triangle, we are somehow (directly?) aware of Plato's Triangle. The second answer cannot account for why on a particular occasion we should experience a triangle and not a circle. There would be nothing special about the physical relations between the triangle selected and some other piece of physical space with a different geometry, to account for why it should be the triangle and not the square known to exist. Furthermore, we could not account for why it should be one triangular region of space rather than another.

Unless that subspace has a causally relevant boundary. (I meant the latter answer to exclude this possibility.) This is the only possible answer, I think. The model for knowledge of hallucinatory experience that is the most plausible is that the experience, event, property, what have you, takes place, and this causes

---

[8] The question can be rephrased but still asked by those who deny the instantiation of afterimage and hallucination predicates: when we allegedly afterimage, we are in a certain relation to something with a certain geometry; can the latter thing be identified with anything physical?

in us the belief that it takes place. For its particular geometrical nature to be known requires that the geometrical properties be causally operative, and cause different beliefs for different geometrical shapes. This means that at least the boundary of the triangle be causally operative in introspection and this would seem to mean that the triangular thing be made up of something, some stuff (this is not to deny that it be made up of collections of cells, or alternatively some dualist stuff, or something else). A patch of space simply singled out in thought from surrounding patches would not give us sufficient causal distinctiveness.

There would seem to be only one reasonable possibility if physicalism is to be true, and that is that the triangle is a part of the body, presumably the central nervous system. (See below, section 10, for an alternative.)

So, then, for physicalism to be true, we will need to be able to detect distinctive pieces of brain tissue of the right shape whenever we afterimage or hallucinate a given geometrical figure. (Later we will look at the possibility that the items of the right shape be electromagnetic.)

What a crude demand on physicalism ! This demand on physicalism is one that has been overwhelmingly rejected by physicalists. To this I can only reply that this is where our argument has led us.

But surely, it might be replied, sensing machines can be constructed whose internal configurations need bear no geometrical relationship to the items whose geometry they are distinguishing. We have already dealt with that argument. It is not being claimed that experience

must be the way I am saying it is. I see no reason to deny that the Direct Realist machine might be constructed one day. I am claiming that this is how experience is for humans.

9. The Topology of Neural States.

We now turn to look at neurophysiology, to see whether our "crude" conclusion really is so crude. We immediately find that one school of psychology thought precisely as I have - Gestalt psychology. Their position seems largely to have been rejected by more modern physiologists (c.f. Luria 1973 p.229, Pribram 1971 p.468), but, as we will see, it is not quite so obvious that it should be.

The account of the organisation of the human brain that I will present is preceded by a caveat.[9] I base the account principally on the conclusions of two neuroscientists, Luria and Pribram, but it must be stressed that their conclusions are somewhat speculative. Little is known for sure, for instance, about what happens along the causal chain beyond the primary and secondary visual zones of the cortex. Needless to say, visually acquired information is causally relevant to the higher functions of the brain represented further along that chain. Thus, because the conclusion of this thesis depends on the interpretation of neurophysiological evidence given here, that conclusion must be regarded as tentative.

---

[9] I am indebted to Dr. Chris Cooper, Psychology Dept., University of Adelaide, for valuable discussions on the matters in this section. It was he who was responsible for emphasizing to me the tentative nature of the evidence cited in favour of my argument.

When ordinarily seeing a triangle, there is a triangular patch at least somewhere in the head, namely at the retina. (The same is true for afterimaging a triangle.) A triangular patch of firings of receptor cells takes place which is transmitted by bipolar cells to ganglion cells in the optic nerve. Here, and in what follows, when I say that there is a triangular patch of excitation, I mean that the cells which are principally, or mostly, firing are set out in the shape of a triangle. Obviously there are many cells nearby which are not firing so much. Obviously, too, we have to talk about the statistical frequency of the cells firing, for a cell fires discontinuously, on and off. This, together with the fact that the triangular patch at the retina and elsewhere has fuzzy edges, could presumably be accounted for by the physicalist by saying that the consciousness, i.e. the introspective belief "monitor", does not pick up the discontinuities or the firings of the cells around the edges of the triangle which are firing more than their unstimulated neighbours but less than the cells right on the perimeter of the triangle. These points are intended to apply as well to the other triangles in the head which we will show to be present.

It is unlikely, however, that this triangle at the retina is sufficient to account for what is common to seeing, afterimaging and hallucinating a triangle. While seeing and afterimaging are closely associated with retinal events, other sorts of "imaging" (psychologist's term) do not seem to be. For instance, hallucinations can be induced by direct electrical stimulus of the visual cortex (Luria Ch. 3, 8), which is some way along the causal path from, and spatially distinct from, the retina.

The two optic nerves, one from each eye, converge at the optic chiasm, and then diverge to different hemispheres of the brain. In the centre of the retina is a spot of maximum colour sensitivity, the fovea, and it is important to note that images to the left of the fovea in both eyes, are sent to the left hemisphere of the brain. Similarly, the right half of both retinas is represented (initially, at least) in the right hemisphere.

The first point of call for the impulses after the chiasm are the left and right lateral geniculate bodies (LGB). Cells in the lateral geniculate body, due to their various connections with retinal cells, are sensitive to retinal simulation roughly in the form of a circle.[10] (These are not the only cells in the LGB, of course. The picture we are presenting is an extremely simplified one.) The important points about the LGB for our purposes are twofold. First, influences at neighbouring points on the retina are transmitted to spatially neighbouring points in the LGB (with an important proviso, to be mentioned later). In topological terms, this amounts to neighbourhoods of a point being mapped to neighbourhoods of the corresponding point, and so open sets being mapped onto open sets. This is the condition of topological similarity i.e. homeomorphism. Thus a triangle is mapped to a homeomorphic image of a triangle. Our "crude demand" is perhaps not so crude after all. The second point is this: that

---

[10] This is a slight oversimplification. See e.g. Hubel 1963, Hubel & Wiesel 1962, 1963. For a clear exposition of these and other matters in this brief account, see Lindsay Norman 1974, or Cornsweet 1970.

our story has made a certain oversimplification. Recall that impulses from the left hand sides of the retinas go to the left LGB, and similarly for the right hand sides. This means that it is only figures either wholly to the left of the fovea, or wholly to the right, that receive a homeomorphic representation in the LGB. A triangle right at the centre of the visual field, spanning the fovea, will in fact have its left half represented in the left LGB, and its right half represented in the right LGB, with not a single whole triangle occurring in either.

Leaving aside the complication of the divided image, this neighbourhood-to-neighbourhood mapping is preserved at the point of entry to the cortex, too.[11] Brodmann's area 17, as it is called, is also termed the primary visual cortex. It is the only part of the cortex the total excision of which leads to blindness. Areas 18 and 19, the other areas principally concerned with vision, are connected in function with organising the material from area 17 into more complex wholes. To quote Luria's interpretation:

> It follows that the secondary zones of the visual cortex with their complex structure and their facility for the extensive spread of excitation, play the role of synthesising visual stimuli, coding them, and forming them into complex systems. It can therefore be concluded that the function of the secondary zones of the occipital cortex is to convert the somatopical projection of incoming visual excitation into its functional organisation.[12]

Lesions of the secondary zones, for instance, do not lead to blindness or partial blindness.

---

[11] Luria 1973 p.109.

[12] Luria 1973 p.115.

> A patient with a lesion of the secondary visual zones is not blind; he can still see the individual features and, sometimes, the individual parts of objects. His defect is that he cannot combine these features into complete forms, and he is therefore compelled to deduce the meaning of the image which he perceives by drawing conclusions from individual details and by carrying out intensive work where a normal subject perceives the whole form immediately. This can be expressed by saying that the perception of complex visual objects by such a patient begins to resemble the situation in which an archaeologist is attempting to decode a text in unfamiliar script; he readily understands the menaing of each sign although the meaning of the whole text remains unknown. That is why disturbances of visual perception arising from lesions of the secondary visual cortex are not associated clinically with disturbances of the visual field or visual acuity, but are described by the term visual agnosia.[13]

Lesions of the primary zone, on the other hand, lead to blind spots in the visual field, to the point where total excision leads to blindness. These facts are extremely useful in the diagnosis of the location of lesions in the cortex, since the point-to-point projection means that blind spots or areas in the visual field are associated quite closely with corresponding lesions in area 17. Similarly, stimulation of areas of area 17 leads to simple visual hallucinations - flashes, coloured points, etc., in fairly predictable parts of the visual field. Stimulation of areas 18 and 19, on the other hand, has quite a different effect; it differs from that of area 17 with respect to the complexity of hallucinations produced; e.g. flowers and human figures rather than simple shapes.

Other areas of the brain are connected with more complex synthesis of information, unifying various

---

[13] Luria 1973 p.116.

sensory modes etc. It seems, then, though this is at best a tentative conclusion, that if we we are to seek the triangle common to vision, afterimaging, and hallucinating a triangle, we should seek it in area 17.[14]

There are three considerations, of increasing strength, which suggest that the "triangle" in area 17 is such that its triangularity is causally fairly irrelevant to the awareness of the triangle which we have deduced takes place.

The first consideration is that, as in the LGB, area 17 in the left hemisphere represents only the left half of the visual field, and correspondingly for the right hemisphere. A triangle in the centre of the visual field, therefore, does not have a connected representation.

To this point it might be replied by the physicalist that we are simply unaware that our phenomenal visual field is in two halves. When we hallucinate a triagle in the centre of the visual field, we are all but in a certain relation to a triangle. There is not really a whole connected triangle in existence, as we take it to be. Careful introspection just lets us down in this case. But there are two halves of the triangle in existence, and this state of affairs fails to cause in us the belief that there are two halves. We are so constructed that our introspective mechanism "puts together" the two halves of the triangle, in that it forms a belief that there is a single triangle in existence.

---

[14] Areas 18, 19 are associated with recognition of complex patterns as having a certain organisation. But it seems that complex patterns themselves are represented in area 17.

To the extent that such a reply denies the existence of the triangle which we can discern introspectively, so to that extent is the reply counterintuitive. It is perhaps not very counterintuitive.

The second consideration raises something that has been so far skirted over. In area 17 we find a homeomorphic representation of a triangle, true. But, as is well-known, a triangle maps homeomorphically onto a large number of regions of the complex plane. What is in area 17 is in fact more or less elliptical, and certainly well away from being triangular.

In line with the previous reply, the physicalist would have to say that we falsely believe that there is a triangle in existence, and that it is a mistake to believe that a triangle causes our introspective belief that there is a triangle. What happens, the reply presumably goes, is that the elliptical pattern of firings causes by a complex causal route the belief that there is a triangle. To the extent that it does, we can say that we take the ellipse for a triangle, or some such formula.

But if the previous reply allowed us to say that there is nearly a triangle, and it is just a matter of putting the two halves together, this reply commits us to saying that we are quite wrong about the shape of whatever it is that is suitably causally relevant to the production of our belief in the triangle (even supposing that we could make out a suitable causal chain between the ellipse and the belief). If we are that wrong, it is hard to see what interest there is in there being an ellipse there at all. If introspection is systematically so wrong, then it seems to me that we have severed the

connection between the ellipse and the belief enough to be warranted in saying that it is false that we take the ellipse for a triangle. All that happens is that we have caused in us the false belief that there is a triangle. But then the state in area 17 might have been anything you wish, with certainly no necessity for any spatial organisation of firings corresponding to the spatial organisation of retinal stimuli. What I am trying to say is that unless the patch in the cortex is pretty close to being a triangle, so that we can say that we take it for a triangle, we might as well deny any efficacy to introspection for determining the geometric nature of various of our states. If we grant that in hallucinating a triangle introspection enables us to know that we are in a certain relation to a more-or-less triangular thing, then not any deviation from the triangular will do in candidates for the triangular thing. It should not be too far: <u>perhaps</u> splitting in half is allowable, but a big distortion is not.

The third consideration is a more technical one. It is very forcefully put by Pribram[15], and we include a longish extract as the best way of making the point.

> From thalamus to cortex the reverse of the retinal situation holds: a single geniculate cell may contact 5000 cortical neurons, each of which is in contact with some 4000 others through its dendritic fields. This arrangement, aided by inhibitory interactions, insures that, despite some overlap, when two points in the retinal fovea of the monkey are stimulated clear separation is maintained so that two minutes of retinal arc are separated at the cortical surface by 1mm. (Marshall and Talbot, 1941 p.134.) One would think such an arrangement to be compatible

---

[15] Pribram 1970.

with projecting some sort of "image" from the receptor surface onto the cortical surface much as a photographic image is projected onto the film plane surface in a camera.

The paradox appears when the input systems become damaged, either through disease or surgery. True, as expected, a hole (scotoma) can, under the appropriate circumstances, be demonstrated in the visual field in the location predicted from the anatomical arrangement (Fig. 7 - 2). Yet with even the smallest part of the input mechanism intact, this hole is often unperceived even with the eyes held stationary, and pattern recognition, in many respects indistinguishable from normal, remains possible. People with huge scotomata either are wholly unaware of them or can soon learn to get about easily by ignoring them. An animal in whom 80 to 90 per cent to the input mechanism has been removed or interupted is able to solve problems requiring discriminations of patterns differing only in detail. Lashley (1929) removed 80 -90 per cent of the striate cortex of rats without impairing their ability to discriminate patterns. Robert Galambos cut up to 98 per cent of the optic tract of cats and the animals could still perform skillfully on tests necessitating the differentiation of highly similar figures. (Galambos, Norton and Frommer 1967). In a recent experiment, Kao Liang Chow (1970) also working with cats, severed more than three-fourths of the optic tract and removed more than three-fourths of the visual cortex; hardly any of the point-to-point projection system remained intact. Although visual discrimination of patterns became disturbed initially by such drastic interference, the animals relearned the task in about the same number of trials required to learn prior to surgery.

In my experience both in clinical neurosurgery and in the laboratory (e.g. Wilson and Mishkin, 1959), limited removals restricted to cortex that do not massively invade white matter leave the patient or experimental subject's perceptual abilities remarkably intact over the long range. After a temporary scotoma lasting a few weeks, very little in the way of deficit can be picked up.

As already noted, a variety of other methods for disturbing the presumed organisation of the input systems have been tried to no avail: Roger Sperry

and his group (1955) surgically cross-hatched a sensory receiving area and even placed mica strips into the resulting brain troughs in order to electrically insulate small squares of tissue from one another. Lashley, Chow and Semmes (1951) tried to short-circuit the electrical activity of the brain by placing strips of gold foil over the receiving areas. And I have produced multiple punctate foci of epileptiform discharge within a receiving area of the cortex by injecting minute amounts of aluminium hydroxide cream (Kraft, Obrist, and Pribram, 1960 ; Stamm and Pribram, 1961; Stamm and Warren, 1961). Such multiple foci, although they markedly retard the learning of a pattern discrimination, do not interfere with its execution once it has been learned (whether learning occurs before or after the multiple lesions are made). These results make it clear that the effects of sensory input on brain tissue, the input information, must become distributed over the extent of the input system.

Electrical recording has also contributed substantially to the evidence that information becomes distributed in the brain. E. Roy John (John, Hehrington and Sutton, 1967) for instance, uses the technique of "labeling" an input to the visual system by presenting cats stimuli which are differentiated not only by their geometric pattern but also by the frequency of the flickering light which illuminates them. This differential frequency of illumination becomes reflected in the neuroelectric activity of the brain which follows the imposed frequency (or if this is fairly rapid, a subharmonic of that frequency). Thus the frequency encoded difference can be "traced" within the brain. This technique has yielded a number of interesting results, but of importance here is that careful analysis of the labeled wave shapes (computing possible differences between those occurring in one location in the brain and those occurring in others) shows that identical labeled wave forms occur in many brain structures simultaneously.

Another set of experiments performed in my laboratory (Pribram, Spinelli and Kamback, 1967; Figs. 7-3, 7-4) shows, however, that once learning has occurred this distribution of information does not involve every locus within a system. Very small electrodes were used. Monkeys were trained to respond differently to different geometric stimuli. In

contrast to John's experiments, a very brief single flash illuminated the stimuli. Several distinct types of wave forms of electrical activity were evoked in the visual cortex. One type, obtained when the wave form was computed from the moment of stimulus onset, showed clear distinctions that were related to the stimuli. The other two types were obtained when the wave form was computed from the moment of response. One of these reflected whether the monkey received a pellet for responding correctly or whether he did not because he responded erroneously. The other type of wave form occurred immediately prior to the overt response. This wave form correlated with the particular response (pressing a right or left panel of a pair) which followed and was independent of the stimulus shown and the reward obtained. Important here is the fact that all of these characteristic wave forms did not appear everywhere in the visual cortex. One characteristic wave form was recorded from some electrodes, another wave form from other electrodes. Their distribution followed no discernable pattern. However, there was complete consistency from day-to-day and week-to-week of the recordings obtained from any particular electrode. Whatever encording process has occurred, it had stabilised by the time of our recordings.

These experimental results are incompatible with a view that a photographic-like image becomes projected onto the cortical surface. The results do indicate that each sensory system functions with a good deal of reserve. Since it seems to make little difference to overall performance which part of the system is destroyed and which remains, this reserve must be distributed in the system - the stored information necessary to make a discrimination is paralleled, reduplicated over many locations. It thus becomes likely that the retardation in learning resulting from the epileptic foci produced by aluminium hydroxide cream implantations indicates interference with this reduplication of information storage (Fig. 7-5).

The questions raised by these observations must be juxtaposed against another: how do objects appear sufficiently consistent so that we can recognise them as the same, independent of our angle of view or their distance from us? How do we recognise an object regardless of the part of the retina, and

therefore of the brain, which is directly excited by the light coming from that object? The capacity for such size and object constancy is already developed in the human infant a few weeks of age. Thus any easy explanation of the constancy of the phenomenon in terms of learning is brought into question. Just what sort of mechanism would simultaneously allow for the existential flexibility of perception and the constancy of recognition once distribution has taken place?

> Both the facts of pattern perception in the presence of scotomata and of perceptual constancy demand that there must be an effective neurological mechanism to spatially distribute the information contained in the input to the brain. If the facts of perception are to be accounted for, the simple correspondence of a point-to-point ikonic isomorphism suggested by the anatomy of the system cannot be sufficient. When 80 per cent of the visual field is blinded by cortical removal, recognition is mediated by the remainder of the visual field; when the visual cortex is peppered by lesions, the part between the lesions functions so well that little difficulty is experienced in making discriminations; whether we view an object with one part of our retina or another, or whether we view it from one angle or another, we can still recognise the object. These are not the properties of ordinary photographic images - tear off 98 per cent or even 80 per cent of most photographs and try to identify them ! [16]

There is a slight inconsistency in this: when Pribram says that the results "... are incompatible with the view that a photographic-like image becomes projected onto the cortical surface" he does not really mean to deny that there is a point-to-point projection. Pribram's argument is rather that the point-to-point projection can be seriously disturbed without affecting our ability to discriminate shapes, and thus that the point-to-point projection is causally of little relevance to the functioning of the mechanism of perception.

---

[16] Pribram 1970 pp.119-124.

A similar position is defended by Mucciolo (Mucciolo 1974). The severing of large parts of the optic tracts and destruction of large amounts of the visual cortex of rats, cats, etc., by Lashley and Galambos suggests strongly that the physical configurations of firings ~~are~~ not especially causally relevant to experiencing the geometry of our mental states. Pribram's experiment involving inserting large numbers of irritative metal spots into monkeys was explicitly designed to destroy as much as possible the geometrical configurations of firings in the cortex. Large lesions in humans similarly fail significantly to diminish significantly pattern recognition. Mucciolo concludes

> It has been suggested by many defenders of IT that psychological states are identical with the stimulation of certain parts of our brain. Clearly, this approach cannot be reconciled with the kinds of experimental evidence described above.[17]

If, then, the theory of perception ~~that~~ for which we have been arguing ~~for~~ here is correct, and so perception of a triangle necessarily involves there being an "inner" triangle, then it would seem that this inner triangle is not identical with the projection of the retina onto area 17 of the cortex. For that inner triangle is necessary to perception as we know it: take it away, change it to a square, destroy 98 per cent of it, and we no longer have ordinary perception of a triangle. Nor do we have afterimaging a triangle, hallucinating a triangle. Pribram's argument, then, reinforces the conclusion of our two previous arguments. I conclude that there is some reason to believe that the triangle we are aware

---

[17] Mucciolo 1974 p.331.

of in perception, hallucination, etc., is not an arrangement of cells in the human head.

10. Coding Visual Information.

Pribram goes on to suggest an analogy with the hologram in an attempt to explain the ability of the visual cortex to be little affected in its ability to carry information by destruction of widespread and more-or-less arbitrary parts of it. Mucciolo also defends the theory that our mental states might be physically represented by some sort of standing wave-pattern of potentials in the way that a hologram represents the information it contains.

Now there is a general point to be made about any suggested mechanism which attempts to account for mental states in terms of "holistic" wave/field properties of the cortex. (c.f. also Lashley's "aggregate field" theory. See Lashley 1960, Mucciolo 1974). Beloff[18] makes the point that there are (at least) two ways of transmitting and representing information: (1) iconographically i.e. by transmitting something which is like the state-of-affairs to be informed about in those respects it is desired to communicate; and (2) in a code form i.e. by something which is not like the original in those respects, but has corresponding respects of its own, (which can either naturally correspond, or be conventionally decided upon to correspond). Now in either case, to obtain the information, it must be extracted from whatever stores it or is carrying it i.e. we must actually

---

[18] Beloff 1962 p.73.

come to beliefs corresponding to the facts being communicated. This being so, the second system, the coding system, suits a Direct Realist picture of perception very well. Events occur, are encoded in light rays and transmitted to the retina, where they are coded again into electrical impulses which are eventually decoded as beliefs. The decoding is not a conscious or rational process, any more than it is in a decoding machine. There is simply some mechanism with electrical inputs and belief outputs. (This, albeit dimly, is what Lord Brain seems to be driving at when he uses the concept of "information" in an attempt to explain perception.[19]) Pribram's hologram model, similarly, functions to account for the redundancy of information in the visual cortex, but it is essentially a system for storing and transmitting information. A hologram does not store geometrical information iconographically. To extract geometrical information from a hologram requires that something be done to it. The triangle is ~~only~~ there in the standing wave pattern only in the ~~same~~ way that certain operations performed on a hologram enable the triangle to be reproduced in reflected light. The extraction of information, that is the formation of beliefs, is something additional.

If such a system of coding suits Direct Realism, it does not suit the picture of perception (and certain perceptual aberrations) that we have argued for. Causally after the retinal event, and causally before the formation of beliefs about the world, there occurs an event which is not merely a coded representation of

---

[19] e.g. Brain 1966 Chapters 2, 3.

what, under normal circumstances, causes it. It has some of the features of an iconographic representation, specifically certain geometric ones. Evidently, any causal account should therefore have to be modified from the previous (Direct Realist) one, to include a partial decoding of information into something with the requisite iconographic features. But, as we have just concluded, there does not seem to be anything in the brain or in a holographic pattern with both the suitable geometric, and suitable causal properties, to identify this iconographic item with.

We have argued principally about the geometry of our perceptual and quasi-perceptual states. We have avoided looking at the physiological facts concerning colour. The principle reason for this is that, as I understand it, there is little that is known for sure about colour mechanisms in the cortex which is coherent enough to contribute to this debate. Let us just observe that if in perception there really is involved a red* triangle, wholly red*, so that its parts are red* too, and if that triangle is neither in the body, nor anywhere else in physical space, nor in Plato's heaven, then it is unlikely that its redness* can be identified with anything in the body either.

To repeat a point made previously: the conclusion of this section is very tentative. The conclusion is that what we know of the brain <u>suggests</u> that such causal mechanisms as are involved do not jointly have the properties of correct topological structure and sufficient causal relevance to make them identifiable with the geometrical items present in perception. The structures proposed by Pribram and Mucciolo to have

adequate causal relevance have irrelevant geometry. The structures with something approaching relevant geometry seem to be causally relatively of little moment. The conclusion of this thesis, then, is weak. It is only that there is some reason to think that physicalism is false.

The foregoing argument cuts across the previously mentioned defences to physicalism: topic neutral analysis, elimination, Ramsey Sentences, adverbialisation and property identification. For the remainder of this chapter we will look at the relationship between the argument and those defences.

11. Defences of Physicalism.

First, let us look at property identification. Our approach to property identification attempted to allow as much leeway as possible for someone who wanted to identify the various properties we have been discussing with neural properties. Certainly we were prepared to permit contingent property identifications. In the case where we might wish to make a property identification via the reduction of co-extensive predicates, we offered an account of where it is that co-extensive predicates denote identical properties.

But not every pair of property names refer to the same property. We obviously cannot allow total licence in property identifications. How should we decide if the properties we were concerned with are identical, or, alternatively, non-identical with certain other sorts of properties? One side of this decision has a certain presumption in favour of it: Ockham's Razor and like Principles. Therefore, we should hold

that the identification is to be made unless positive reason can be given to deny it.

What sort of positive reason could there be? Surely only a reason which derives from the considerations that lead us to isolate the troublesome properties in the first place; a reason deriving from whatever it was that made us inclined to think that the properties were troublesome and hence were instantiated. It would seem that such a reason must derive from that faculty we have for knowing certain things about ourselves without (or without apparently) going through the normal sensory channels.

So we were led to look at just what we could reasonably say about ourselves on those occasions when we seem to exercise that faculty. We saw that we were able to conduct our investigation without presuming that certain predicates e.g. "x has a red afterimage" were instantiated. We found, though, that on those occasions, we (or our states) possessed complex properties which were not to be identified with certain physicalistically acceptable properties: namely similarities, differences, causal properties and suppressed tendencies to have beliefs.

In saying "found", I presuppose that what in the above is supposed to have been found, is true. At any point in all this, the philosopher who denies incorrigibility is at liberty to deny that such claims are in fact true: we were able to argue without presupposing that we had afterimages; but we were not able to argue without concluding that we did know certain things about ourselves. We invented terms to describe our states and properties so as to avoid presupposing against the

eliminationist that we already had words for them.

The final upshot of our investigation was that on the disputed occasions, we were in a certain relation to a certain entity with a certain geometry. If we are to be wholly physical creatures, then, it should be possible to discover that figure somewhere in the physical universe. To put it slightly differently, the complex properties we discovered about ourselves, and which we wished to identify with physical properties, turned out to have a relational component. The property identification, then, should map this relation into an acceptably physical relation, and an obvious necessary condition for relation identification, is identity of the terms of the relation. But we can find nothing physical to identify our geometrically shaped objects with.

So the strategy I have employed is this: to claim that we can come to know enough about what we are like on certain occasions, that is, what properties we have on those occasions, to prevent those properties being identified with physical properties.

If the argument of this chapter is correct, it prevents any such property identification, contingent or not. As we have already seen, there is something of a problem about deciding whether what we would have had would have been a contingent identification. But it follows from our conclusion that a non-contingent identification of the properties cannot be made either. Now it seems clear that it would be sufficient to make a non-contingent identification of the properties, if we could succeed with a topic neutral analysis of the troublesome predicates. But here we must be clear that we mean <u>all</u> the troublesome predicates. We might dispose

of the extant troublesome predicates only to find that when we examine the matter more closely we find that we wish to say more about ourselves and our states than topic neutral predicates give the power to say.

If we invent new words to describe ourselves, whatever our motives might be, then we do not so clearly allow the possibility of a topic neutral analysis. What an analysis might do for us, though, is to "analyse away" the content of what it is being claimed we know about ourselves. To that move, and to eliminationist and adverbialist moves, it can reasonably be said that if successful they would have the function of denying the truth of what I am claiming we know. If I am right, then, they cannot work. I am suggesting, further, that careful introspection constitutes part of the test of whether a given piece of analysis, or elimination, works. It is not easy to disregard careful introspection, at least with respect to such obvious things as (alleged) pains and (alleged) afterimages. It is surely unreasonable to disregard it solely on the grounds that accepting it leads to dualism.

In particular, it seems to me that the situation is the same for Ramsey Sentence theorists. The principle innovative virtue of the Ramsey Sentence approach is that it sees that in order to give the meaning of a term it might be necessary to teach a person a whole theory, that some terms only take on their meaning when with other terms they function in a theory, and that thus the meaning of the term is given by the place that it occupies in the theory. The virtue of this is that it makes it at least prima facie possible to claim that anything looking like a traditional analysis

of mental predicates is not on, while at the same time holding our for some of the advantages of topic neutrality. It is common-sense psychology which can be said to be topic neutral.

But as we have already said, the approach still rests on the assumption that common-sense psychology is topic neutral. (If not, if it were woefully dualistic, what could one do but fall back on elimination or property identification?) And surely this assumption can be met by asking what common-sense psychology is conceived to be. I do not mean now to be alluding to the differences that we found between Smart's conception of it and Lewis'. I mean rather: is common-sense psychology just those facts which people mostly agree on, or is it permitted to include whatever we can find out about the mind by introspecting and deducing? If the first, you beg the question against the possibility of the second disjunct. If the second, then you have to submit to the investigation we have undertaken in this chapter. If we are correct in our conclusion about that investigation, then common-sense psychology, while perhaps not <u>woefully</u> dualistic is certainly <u>ineradicably</u> so, because the dualist bits we have found are among the true bits. It is no use to say that a near realisation of common-sense psychology might be physicalist. So it might be, but there is more than one way for there to be a near realisation of a theory. One way is if certain parts of the theory are false, and in particular if we do not need to find extensions for certain predicates in the theory for they are not in actuality instantiated. That is the way that Lewis obviously had in mind. But another way might be if

common-sense psychology were just a little dualist but those dualist predicates were really instantiated, so that the physicalist near realisation actually left out something really there. (This reply concedes only for the sake of argument that the mind is only a little bit dualist.) Again, we have that reliance on Principles of scientific Method to make reasonable the belief in the future discovery of the truth of the Identity Theory, only goes through in the absence of a conclusive argument for the opposite conclusion. Where could such an argument come from? Smart and Lewis do not address themselves to that question, but it seems that if it could come from anywhere it should come from the place where all the fuss starts: introspection.

12. Dualism.

I have said above that dualism is true. I need not have said that, but only that physicalism seems to be false. Saying it provokes the response: what is dualism? What sort of objects are you telling us exist? How can we begin to evaluate your theory unless you can tell us the nature of the objects in it? It is not unusual to hear Identity theorists demanding of their opponents that they produce their theory. The aim, of course, is to range the candidate theories alongside one another for purposes of comparison by such criteria as counterintuitiveness, simplicity, etc.

The situation here is a little different. I claim to have an argument against one class of theories. It is pointless offering your theory for comparison with a lot of false theories to see which of them is true. However, it would be disingenuous to give this reply as

your sole reason for avoiding producing your theory if you had one. So now the cat is out of the bag: I do not have an alternative theory of the nature of the problematic properties.

I suggest, however, that this situation affords us a prime opportunity for further research. We can discern enough about the problematic properties to conclude that they are not identical with certain physical properties. That is not very much. Let us try, then, to investigate the problematic properties further in order to learn more about their nature. Are they somehow generated by brain fields similar to those suggested in another connection by Lyall Watson?[20] Perhaps. Just because we cannot locate the bearers of certain of the properties (specifically the topological ones) in the physical universe, it does not mean that we cannot investigate those properties by means other than introspection - even if all methods of investigation had to contain an element of introspection. It is surely wrong to think that if we could find out enough about entities of a certain sort, and they turned out to be causally well-behaved, we would have to count them physical. None of the supernaturalist, occultist or religious stories represent the Other World as chaotic. If entities are causally well-behaved, and their interactions with the physical are equally lawlike, this does not prevent those entities from being quite queer nonetheless: emergent, nonspatial, surviving death perhaps. This being so, dualism should be regarded as a research programme. Indeed, I think that this is the best way to see dualism. After all, we have no systematic account of the nature and laws of entities

---

[20] Watson 1973.

other than those in space made up of waves and particles. (We do not have a final theory of the latter, either, but at least we have made a start.) If dualism is to be anything more than a denial of physicalism, then, it must be seen as programmatic.

A final point: the argument given in this book, if correct, seems to establish that physicalism is contingently false. What then would the mind be like if physicalism were true? One possibility harks back to what we said about perception: if Direct Realism were true, perception would just be beliefs, aberrant perception would be misplaced beliefs, introspection would be beliefs about beliefs, the latter perhaps not seeming to be beliefs. Another possibility might be that perception is mediated by non-belief states which can be known directly to be physical. In introspection it might seem to us that we were a brain, being affected in various ways by stimuli the nature of which could be established conclusively only by hypothesis. A variant of the last is where we do not know our mediating sensations to be physical, but then we do not know anything about them that prevents their being physical either. I do not think that I can describe this last possibility any more precisely. These last two possibilities are, admittedly, rather fanciful. But then it stands to reason that it would be difficult to imagine how a radically different mechanism for experience would seem to the experiencer.

APPENDIX

Some Improved Definitions

The purpose of this appendix is to improve on certain of the definitions given in Part One, in order to remove some problems of exposition. In Chapter One, the simplifying assumption that we will one day be in possession of the final theory was introduced. There is a problem connected with this assumption. It is, that if we permit nonsynonymous predicates to express the same property, then there is no reason to think that if a predicate α nonsynonymously expresses the same property as some physical predicate β (i.e. $\beta \varepsilon P$) which belongs to the language of the final theory, then α will not occur in the language of the final theory. To say that α need not occur in a theory does not entail that it will not occur. Moreover, if α expresses a physical property, then surely there can be no harm in including α in the final theory. But if we permit the possibility that such an α occur in the final theory, then, if we remain with our definition of physicalism, contingent property identification cannot be seen as part of an eliminative materialist's methodology, contrary to what was claimed in Chapter Six. However, it seems to me that there are unificatory advantages in making this latter claim. I propose, therefore, the following definitions. If they are adequate, then the claim can be maintained.

First, we must dispense with the assumption that the final theory will one day be discovered. We must fall back on the notion that there can be a theory which is never exhibited, and that two theories can be distinct even though both are never exhibited or discovered. If we wish to remain with the idea that a theory is a set of items (e.g. sentences) closed under deducibility - and it seems to me that this is correct, and furthermore that too much in this thesis hangs on holding it - then we will have to say that the members of a theory are more like propositions. This in turn might introduce the tension of talking about the parts of propositions, for there are obvious advantages in being able to speak of predicates' being members of the language of a theory. In order to solve this, I will continue to say that a theory is a set of sentences, where sentences are constructed from predicates, quantifiers, and truth functional operators, but conceive of sentences as having non-extensional identity conditions, and leave the problem (which is hardly a new one) of what those identity conditions are, unsolved.

Then we can say that a theory is a <u>final</u> theory, if it is true and complete i.e. if it is true and every fact about which things exist and which properties and relations they have is stated in it. <u>Physicalism</u> is the doctrine that at least one final theory of the universe is a physical theory i.e. one all of whose predicates come from P (where the membership of P is as indicated in Chapter One: P contains only the predicates of physics and inorganic chemistry).

I mention a problem about these definitions. If sentences exist, then a final theory should be able

to deal with properties like their truth, or the satisfiability of their predicates. Following Tarski, this would have to be done by stratifying the theory and its language, but the introduction of primitive semantic notions might be a difficulty with a set P containing only predicates from physics (including set theory) and chemistry. I will not pursue this problem.

Finally we arrive at the definition of eliminability. A predicate α, is eliminable if there is a final theory in whose language neither α nor any synonym of α occurs.

This definition of eliminability enables us to subsume both Rorty's position and the position of the physicalist contingent-property-identificationist under eliminative materialism.

First, a clarification point. As was noted in Chapter Five, it is not clear that the falsity of a sentence of the form "$(Ex)\phi x$" is a sufficient condition for the eliminability of the predicate "ϕx". It might be argued that a theory can contain laws which relate predicates while being neutral on the question of whether anything ever satisfies those predicates. For example, there might be laws relating the energies of systems which allow an infinite range of possible energies to systems (e.g. with no upper limit), and the theory say nothing about whether sufficiently high energies are ever realised in physical systems. Indeed, it might be the case that these energies are never realised, and so the final theory have no need, as regards describing the properties of what exists, for predicates expressing them.

I have two points to make about this. The first point is that I am not convinced that it might not be possible to prove (somewhat like Craig's theorem) that for any theory containing such predicates, there is a theory which is equivalent as to existential consequences, and which does not contain them. (Possible proof: take the subset of existential consequences, and close it under deducibility. Mixed existential and universal quantifiers pose a problem for the definition of "existential consequences".) I suggest this only as a possibility; but if it were true, it would mean that a final theory could be extracted from any such theory. The second point is that it seems to me that it is unlikely that this problem arises in connection with the problem of eliminating mental predicates. It is hard to see what unificatory purpose, or other purpose, might be served by retaining mental predicates if they are never satisfied.

This being so, then we can say that at least for a mental predicate, say "Mx", it is a sufficient condition for its being eliminable, that "(Ex)Mx" is false. It follows from this that, as we have interpreted it, Rorty's position is a version of eliminative materialism.

Now let us turn to property identifications. Suppose that a mental predicate, say "Mx", expresses the same property as some nonsynonymous predicate, say "Bx", which is a member of P. Then there is no need for a final theory to include "Mx" in its vocabulary. From this it does not follow that no final theory contains "Mx". But it does seem reasonable to conclude that some final theory does not contain "Mx". For, select one which does contain "Mx", and replace "Mx" wherever

it occurs by "Bx". Then, plausibly, the new theory states the same facts about which things exist and their properties and relations, and does not have "Mx" in its vocabulary. But, by the definition of eliminability, "Mx" is eliminable if some final theory does not contain it. Therefore, "Mx" is eliminable. Thus, the position of the physicalist-contingent-property-identificationist is a form of eliminative materialism.

BIBLIOGRAPHY

1. Works Referred To In The Text.

1. Achinstein 1974 — P. Achinstein "The Identity of Properties" American Philosophical Quarterly XI(1974), 257-274.

2. Armstrong 1961 — D.M. Armstrong Perception and the Physical World London: Routledge & Kegan Paul, 1961.

3. Armstrong 1968 — D.M. Armstrong A Materialist Theory of the Mind London: Routledge & Kegan Paul, 1968.

4. Armstrong 1973 — D.M. Armstrong Belief, Truth and Knowledge Cambridge: Cambridge University Press, 1973.

5. Ashby 1964 — W.R. Ashby An Introduction to Cybernetics London: Methuen & Co. (University Paperbacks) 1964.

6. Aune 1967 — B. Aune Knowledge, Mind and Nature New York: Random House, 1967.

7. Beloff 1962 — J. Beloff The Existence of Mind London: Macgibbon & Kee, 1962.

8. Bernstein 1968 — R.J. Bernstein "The Challenge of Scientific Materialism International Philosophical Quarterly, VIII, 2 (June 1968), 252-275.

9. Bradley 1963 — M.C. Bradley "Sensations, Brain-Processes and Colours" Australian Journal of Philosophy, XLI, 3 (December 1963), 385-393.

10. Bradley 1964 — M.C. Bradley Critical notice of Philosophy and Scientific Realism, by J.J.C. Smart Australasian Journal of Philosophy, XLII, 2 (August 1964) 262-283.

11. Brain 1966 Lord Brain Science and Man London: Faber and Faber Ltd., 1966.

12. Brandt 1960 R.B. Brandt "Doubts about the Identity Theory" in S. Hook (ed) Dimensions of Mind New York: Collier Books, 1960 62-70.

13. Brandt & Kim 1967 R.B. Brandt and J. Kim "The Logic of the Identity Theory" Journal of Philosophy LXIV, 17 (September 7, 1967), 515-537.

14. Campbell 1970 K. Campbell Body and Mind Macmillan, 1970.

15. Carnap 1963 R. Carnap "Herbert Feigl on Physicalism" in P.A. Schilpp (ed) The Philosophy of Rudolph Carnap La Salle, Illinois: Open Court, London: Cambridge University Press, 1963, 882-886.

16. Causey 1972 R. Causey "Attribute-identities in Micro-reductions" The Journal of Philosophy LXIX, 14 (August 3, 1972) 407-422.

17. Chandler 1970 J.H. Chandler "Incorrigibility and Classification" Australasian Journal of Philosophy XLVIII (1970) 101-106.

18. Clark 1970 R. Clark "Concerning the Logic of Predicate Modifiers" Nous 4 (1970) 311-335.

19. Chisholm 1957 R. Chisholm Perceiving Ithaca, N.Y.: Cornell University Press, 1957.

20. Chisholm 1966 R. Chisholm Theory of Knowledge Prentice-Hall Inc., Englewood Cliffs, N.J., 1966.

21. Chisholm 1970 R. Chisholm "Events and Propositions", Nous 4 (1970), 15-24.

22. Chisholm 1971 R. Chisholm "States of Affairs Again", Nous 5 (1971), 179-189.

23. Cornman 1962 J.W. Cornman "The Identity of Mind and Body" The Journal of Philosophy LIX, 18 (August 30, 1962), 486-492.

24. Cornman 1968 J.W. Cornman "On the Elimination of 'Sensations' and Sensations" The Review of Metaphysics, XXII (1968) 15-35.

25. Cornman 1968a J.W. Cornman "Mental Terms, Theoretical Terms, and Materialism" Philosophy of Science XXXV, 1 (March 1968), 45-63.

26. Cornman 1971 J.W. Cornman Materialism and Sensations New Haven and London: Yale University Press, 1971.

27. Cornman 1972 J.W. Cornman "On Direct Perception" The Review of Metaphysics XXVI(1972).

28. Cornsweet 1970 T. Cornsweet Visual Perception New York: Academic Press 1970.

29. Davidson 1967 D. Davidson "The Logical Form of Action Sentences" in N. Rescher (ed) The Logic of Decision and Action Pittsburgh: University of Pittsburgh Press, 1967, 81-95.

30. Davidson 1970 D. Davidson "Mental Events" in L. Foster and J.M. Swanson (eds) Experience and Theory University of Massachusetts Press, 1970.

31. Davidson 1974 D. Davidson "Belief and the Basis of Meaning" and "Replies to David Lewis and W.V. Quine" Synthese 27 (1974) 309-323 and 345-349.

32. Davis 1958 M. Davis Computability and Unsolvability New York, Toronto, London: The McGraw-Hill Book Company, Inc. 1958.

33. Donellan 1972 See D. Davidson and G. Harman (eds) Semantics of Natural Languages Dortrecht-Holland: D. Reidel Publishing Co., 1972.

34. Ducasse 1951 C.J. Ducasse Nature, Mind and Death La Salle, Illinois: Open Court Publishing Co., 1951.

35. Edwards 1967 P. Edwards (ed) The Encyclopedia of Philosophy Prentice-Hall, Inc., Englewood Cliffs, N.J. 1967.

36. Feigl 1963 H. Feigl "Physicalism, Unity of Science and the Foundations of Psychology" in P.A. Schilpp (ed) The Philosophy of Rudolph Carnap La Salle, Illinois: Open Court, London: Cambridge University Press, 1963, 227-267.

37. Feigl 1967 H. Feigl The 'Mental' and the 'Physical': the Essay and the Postscript Minneapolis: University of Minnesota Press, 1967.

38. Feyerabend 1963 P.K. Feyerabend "Materialism and the Mind-Body Problem" Review of Metaphysics XVII, 1 (September 1963), 227-267.

39. Fodor 1965 J. Fodor "Explanations in Psychology" in M. Black (ed) Philosophy in America Ithaca: Cornell University Press, 1965, 161-179.

40. Fodor 1968 J. Fodor Psychological Explanation Random House, 1968. Pages 90-120 reprinted as "Materialism" in D.M. Rosenthal (ed) Materialism and the Mind-Body Problem Prentice-Hall Inc., Englewood Cliffs, N.J., 1971, 128-149. Page references to the latter.

41. Hirst 1959 R.M. Hirst The Problems of Perception London: George Allen & Unwin Ltd., 1959.

42. Hubel 1963 D. Hubel "The Visual Cortex of the Brain" in Scientific American November 1963.

43. Hubel & Wiesel 1962 D. Hubel and T. Wiesel "Receptive Fields, Binocular Interaction and Functional Architecture in the Cat's Visual Cortex" Journal of Physiology 160 (1962) 106-154.

44. Jackson 1973 F. Jackson "Is there a Good Argument Against the Incorrigibility Thesis?" Australasian Journal of Philosophy LI (1973), 51-62.

45. Jackson 1974 F. Jackson "The Existence of Mental Objects" forthcoming.

46. Kalke 1969 W. Kalke "What is wrong with Fodor and Putnam's Functionalism" Nous 3 (1969) 83-93.

47. Kemeny & Oppenheim 1956 J. Kemeny and P. Oppenheim "On Reduction" Philosophical Studies, 1956, 6-19.

48. Kim 1966 J. Kim "On the Psycho-Physical Identity Theory" American Philosophical Quarterly, III, 3 (July 1966), 227-235.

49. Kim 1968 J. Kim "Reduction, Correspondence and Identity" The Monist , LII, 3 (July 1968), 424-438.

50. Kim 1970-1 J. Kim "Materialism and the Criteria of the Mental" Synthese 24 (1971) 323-345.

51. Kim 1972 J. Kim "Phenomenal Properties, Psychophysical Laws and the Identity Theory" The Monist 56, 2 (April 1972), 177-192.

52. Lashley 1960 K. Lashley "In Search of the Engram" in F. Beach et al (eds) The Neurophysiology of Lashley New York: McGraw Hill, 1960.

53. Lean 1953 M. Lean Sense Perception and Matter London: Routledge and Kegan Paul, 1953.

54. Lewis 1966 D. Lewis "An Argument for the Identity Theory" Journal of Philosophy LXIII (1966) 17-25.

55. Lewis 1970a D. Lewis "Psychophysical and Theoretical Identifications" read to Philosophy Colloquium, University of Huston, April 10, 1970, 1-21. Published in Australasian Journal of Philosophy L (1972) 249-258 . Page references to the former.

56. Lewis 1970b D. Lewis "How to Define Theoretical Terms" Journal of Philosophy LXVII (1970) 427-446.

57. Lewis 1974 D. Lewis "Radical Interpretation" Synthese 23 (1974) 331-344.

58. Lindsay & Norman 1972 D. Lindsay and D. Norman Human Information Processing New York: The Academic Press, 1972.

59. Locke 1967 D. Locke Perception and Our Knowledge of the Physical World London: Allen and Unwin, 1967.

60. Luria 1973 A.R. Luria The Working Brain Penguin Books, Harmondsworth, Middlesex, 1973.

61. Lycan 1974 W. Lycan "Mental States and Putnam's Functionalist Hypothesis " Australasian Journal of Philosophy LII (1974), 48-62.

62. Lycan & Pappas 1972 W. Lycan and G. Pappas "What is Eliminative Materialism" Australasian Journal of Philosophy L (1972) 149-159.

63. Malcolm 1963 N. Malcolm Knowledge and Certainty: Essays and Lectures Prentice Hall Inc., Englewood Cliffs, N.J., 1963.

64. Malinas G. Malinas "Physical Properties" Philosophia, 3, 17-31. Page references are to the typed manuscript.

65. Maund 1974 B. Maund "The Epistemological Objection to the Representative Theory of Perception" read to the Annual Conference of the Australian Association of Philosophy Canberra 1974. Forthcoming.

66. Medlin 1967 B. Medlin "Ryle and the Mechanical Hypothesis" in C.F. Presley (ed) The Identity Theory of Mind, Brisbane: The University of Queensland Press, 1967, 94-150.

67. Meehl & Sellars 1956 P. Meehl and W. Sellars "The Concept of Emergence" in H. Feigl and M. Scriven (eds) Minnesota Studies in the Philosophy of Science Vol 1 Minneapolis, University of Minnesota Press, 1956.

68. Miller 1974-5 D. Miller "The Accuracy of Predictions" Synthese 30 (1975), 159-191.

69. Mucciolo 1974 L. Mucciolo "The Identity Thesis and Neurophysiology" Nous 8 (1974), 327-342.

70. Nagel 1961 E. Nagel The Structure of Science London: Routledge & Kegan Paul, 1961.

71. Oppenheim & Putnam 1958 P. Oppenheim and H. Putnam "Unity of Science as a Working Hypothesis" in H. Feigl, G. Maxwell and M. Scriven (eds) Minnesota Studies in the Philosophy of Science Vol II Minneapolis: University of Minnesota Press, 1958.

72. Pitcher 1971 G. Pitcher A Theory of Perception Princeton University Press, Princeton, N.J., 1971.

73. Place 1956 U.T. Place "Is Consciousness a Brain Process?" British Journal of Psychology XLVII, Part I (February 1956) 44-50.

74. Presley 1967 C.F. Presley (ed) The Identity Theory of Mind Brisbane: The University of Queensland Press, 1967.

75. Pribram 1971 K. Pribram Languages of the Brain Prentice-Hall Inc., Englewood Cliffs, N.J., 1971.

76. Putnam 1960 H. Putnam "Minds and Machines" in S. Hook (ed) Dimensions of Mind London: Collier-Macmillan Ltd., 1960 130-164.

77. Putnam 1967 H. Putnam "Psychological Predicates", reprinted as "The Nature of Mental States" in D. Rosenthal (ed) Materialism and the Mind-Body Problem, Prentice-Hall Inc., Englewood Cliffs, N.J., 1971, 150-161.

78. Putnam 1970 H. Putnam "On Properties" in N. Rescher (ed) Essays in Honour of Carl Hempel New York: Humanities Press, 1970, 235-254.

79. Putnam 1974 H. Putnam "The Refutation of Conventionalism" Nous 8 (1974), 25-40.

80. Putnam "Reductionism and the Nature of Intelligence" unpublished.

81. Quine 1953 W.V. Quine "On Mental Entities" Proceedings of the American Academy of Arts and Sciences, LXXX (1953), 198-203.

82. Quine 1960 W.V.Quine Word and Object Cambridge Massachusetts: The M.I.T. Press, 1960 Page references to the paperback edition, 1964.

83. Quine 1974 W.V. Quine "Comment on Donald Davidson" Synthese 27 (1974), 325-329.

84. Quinton 1955 A. Quinton "The Problem of Perception" Mind LXIV (1955), 28-51.

85. Rennie 1970 M. Rennie "Completeness in the Logic of Predicate Modifiers" Logique et Analyse, 16 (1970) 175-186.

86. Rennie 1974 M. Rennie Some Uses of Type Theory in the Analysis of Language Dept. of Philosophy, Research School of Social Sciences, A.N.U. Monograph Series No.1.

87. Rennie & Malinas 1970 M. Rennie and G. Malinas "The Logic of Predicate Modifiers and its Applications", forthcoming.

88. Rorty 1965 R.Rorty "Mind-Body Identity, Privacy, and Categories" Review of Metaphysics, XIX, 1 (September 1965) 24-54. Reprinted in D. Rosenthal (ed) Materialism and the Mind-Body Problem Prentice-Hall Inc., Englewood Cliffs, N.J., 1971, 174-199, from which page references are taken.

89. Rorty 1970a R. Rorty "In Defence of Eliminative Materialism" Review of Metaphysics, XXIV, 1 (September 1970) 112-121. Reprinted in D. Rosenthal (ed) Materialism and the Mind-Body Problem (see above), from which page references are taken.

90. Rorty 1970b R. Rorty "Incorrigibility as the Mark of the Mental" The Journal of Philosophy LXVII, 12 (June 25, 1970) 399-424.

91. Rorty 1972 R. Rorty "The World Well Lost" The Journal of Philosophy LXIX, 1, (26 Jan. 1972) 649-665.

92. Russell 1959 B. Russell The Problems of Philosophy New York: Oxford University Press, 1959.

93. Scott 1967 D. Scott "Existence and Description in Formal Logic" in R. Schoenman (ed) Bertrand Russell: Philosopher of the Century London: Allen and Unwin, 1967.

94. Sellars 1963 W. Sellars Science, Perception and Reality Routledge and Kegan Paul Ltd. 1963.

95. Sellars 1968 W. Sellars Science and Metaphysics London: Routledge and Kegan Paul Ltd. 1968.

96. Shaffer 1963 J. Shaffer "Mental Events and the Brain" The Journal of Philosophy LX, 6 (March 14, 1963) 160-166.

97. Smart 1959 J.J.C. Smart "Sensations and Brain Processes" Philosophical Review LXVIII (1959) 141-156.

98. Smart 1963a J.J.C. Smart Philosophy and Scientific Realism London: Routledge and Kegan Paul, 1963.

99. Smart 1963b J.J.C. Smart "Materialism" The Journal of Philosophy LX, 22 (Oct. 24, 1963) 651-662.

100. Smart 1967 — J.J.C. Smart "Comments on the Papers" in C.F. Presley (ed) The Identity Theory of Mind Brisbane: The University of Queensland Press, 1967, 84-93.

101. Smart 1970-1 — J.J.C. Smart "Reports of Immediate Experiences" Synthese 22 (1971) 346-359.

102. Verges 1974 — F. Verges "Jackson on Incorrigibility" Australasian Journal of Philosophy LII, 3 (December 1974) 243-250.

103. Watson 1973 — L. Watson Supernature London: Hodder Paperbacks Ltd., 1973.

104. Wittgenstein — L. Wittgenstein Philosophical Investigations Oxford: Basil Blackwell, 1958.

2. Works Referred to by Pribram

1. Kao Lang Chow 1970 — K.L. Chow "Integrative Functions of the Thalamo-cortical Visual System of the Cat" in K. Pribram et al (eds) The Biology of Memory New York: Academic Press, 1970 273-292.

2. John, Herrington & Sutton 1967 — E. John, R. Herrington and S. Sutton "Effects of Visual Form on the Evoked Response" Science 1967, 155, 1439-42.

3. Kraft, Obrist & Pribram 1960 — M. Kraft, W. Obrist and K. Pribram "The Effect of Irritative Lesions of the Striate Cortex on Learning of Visual Discriminations in Monkeys" Journal of Comparative Physiology and Psychology 1963, 50, 17-22.

4. Lashley 1929 K. Lashley Brain Mechanisms and Intelligence Chicago: University of Chicago Press, 1929.

5. Lashley, Chow & Semmes 1951 K. Lashley, K. Chow and J. Semmes "An Examination of the Electrical Field Theory of Cerebral Integration" Psychological Review 1951, 58, 123-136.

6. Marshall & Talbot 1941 W. Marshall and S. Talbot "Recent Evidence for Neural Mechanisms leading to a General Theory of Sensory Acuity" in J. Cattell (ed) Biological Symposia Lancaster The Jacques Cattell Press, 1942, 117-164.

7. Pribram, Spinelli & Kamback 1967 K. Pribram, D. Spinelli and M. Kamback "Electrocortical Correlates of Stimulus Response and Reinforcement" Science 1967, 157, 94-96.

8. Stamm & Pribram 1961 J. Stamm and K. Pribram "Effects of Epileptogenic Lesions of Inferotemporal Cortex on Learning and Retention in Monkeys" Journal of Comparative Physiology and Psychology 1963, 56, 254-260.

9. Stamm and Warren 1961 J. Stamm and A. Warren "Learning and Retention by Monkeys with Epileptogenic Implants in Posterior Parietal Cortex" Epilepsia 1961, 2, 229-242.

10. Wilson & Mishkin 1959 W. Wilson and M. Mishkin "Comparison of the Effects of Inferotemporal and Lateral

10. (cont.) Occipital Lesions on Visually Guided Behaviour in Monkeys" Journal of Comparative Physiology and Psychology 1959, 2, 10-17.