INCORRIGIBILITY AND ELIMINATION: A MENTALIST RESPONSE

by

JOHN ADRIAN BLACK

M.A.(Cantab.) 1975, M.A. Simon Fraser University 1975

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

in
THE FACULTY OF GRADUATE STUDIES
DEPARTMENT OF PHILOSOPHY

We accept this thesis as conforming
to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA
September 1987

© John Adrian Black, 1987
In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of Philosophy
The University of British Columbia
1956 Main Mall
Vancouver, Canada
V6T 1Y3
Date Oct. 22nd 1967
Abstract

Supervisor: Dr. Sam C. Coval

This essay is primarily an examination of a view, propounded by Richard Rorty at the beginning of the last decade, about the nature and existence of minds and mental states. The view is a species of eliminative materialism, and one which is of historical importance in the development of this general position. I argue that it is false. I also attempt to draw some positive conclusions in the philosophy of mind from a criticism of some of its underlying assumptions.

Rorty's fundamental idea is that the belief in the existence of minds and mental states is a primitive scientific theory, which in all likelihood is soon to be overthrown by the superior theory of neurophysiology. It will then be rational, he claims, to deny the existence of minds and mental states. Essential to Rorty's argument for this view is the notion that mental states have a property which the neural states of the replacing theory lack, namely of being the proper subjects of certain incorrigible reports, and which prevents the identification of the two.

I undermine this argument by showing that (i) incorrigibility is not the mark of the mental and (ii) even if
it were, it could not ground the categorial gulf which Rorty sees between mental and physical. I turn then to the major presupposition of the view, that mental states are theoretical entities posited in the causal explanation of behaviour, to see if this characterisation of the mental is an hypothesis adequate to account for the various phenomena of mental discourse. After examining reason-explanation, causal explanation in terms of mental states, the reporting role of mental ascriptions and the non-constative uses of mental language, I find that it is not.

In particular, Rorty's view cannot account for the limited extent to which certain mental reports are incorrigible, nor for the validity of justificatory and non-constative uses of mental language. I argue that the existence of mental states is guaranteed by this validity, and therefore that the issue of their elimination goes beyond considerations of theoretical superiority to the very fabric of human interaction, moral and otherwise. I emerge with the view that ordinary language and neurophysiology are compatible ways of describing people and their behaviour, and that far from being the murky posits of some proto-scientific folk-psychology, mental states are known to exist.
Table of Contents

Abstract ............... ii
Table of Contents ........ iv
List of Figures ........... vi
Acknowledgement .......... vii
Textual Note .............. vii

Chapter 1: General Introduction ........... 1
Footnotes to Chapter 1 .... 12

Chapter 2: Explication of Rorty's View ......... 14
(a) Introduction .......... 14
(b) The Contingent Identity Theory ..... 15
(c) The Demon Analogy ..... 23
(d) The Conceptual Commitments of Mental Language ........ 35
(e) Strategy of the Attack ...... 44
Footnotes to Chapter 2 ...... 47

Chapter 3: Incorrigibility and Incompatibility .... 50
(a) Disappearance versus Non-existence .... 50
(b) Rorty's Form of Incorrigibility .... 53
(c) Traditional Forms of Incorrigibility .... 58
(d) The Criterial Form of Incorrigibility .... 60
Footnotes to Chapter 3 .... 74

Chapter 4: Mental States as Theoretical Entities .... 76
(a) Introduction .......... 76
(b) Actions or Movements? .... 82
(c) Causality and Explanation by Reference to Mental States .......... 89
(d) Actions and Movements (Again) .... 91
(e) Causes and Reasons .... 96
(f) Reasons and Interpretation .... 103
(g) Rationality and Justification .... 107
Footnotes to Chapter 4 .... 111
List of Figures

Figure 1 ................................. 144
Acknowledgement

I am deeply indebted to my supervisor, Sam Coval, for philosophical support throughout the writing of this essay, to committee-members Richard Sikora and Steven Savitt for helpful criticism, to John Westwood, Jonathan Katz, Brian Stoffell, Lawrence Resnick and Jon Young for many conversations on topics relevant to its subject and to the late Robbie Russnaik who helped me develop my view of contemporary psychology.

This essay is dedicated, with fond memories, to Robbie.

Textual Note

Footnotes, marked by numeric superscripts, are collected at the end of each chapter. References to works in the final list of references occur as numerals inside square brackets.
Chapter 1: General Introduction

This essay is a contribution to the dialectic concerning materialism. Its intelligibility may be enhanced if I explain in advance my hopes and wishes as regards the ultimate outcome, or at least the future direction, of this dialectic. It is my opinion that (as a view of Humankind) materialism should be eradicated: I would not however go so far as to suppose that the view is false, since it is not at all clear to me that the differences between materialists and their opponents are in any obvious way disagreements over the truth-values of certain propositions. Nonetheless I believe that the view gives one an incomplete picture of humanity, despite the fact that like any enduring philosophical position it emphasises truths that are worth emphasising, gestures towards intuitions which have somewhere a solid basis.

An incomplete picture of a human is not the picture of an incomplete human: thus I do not suppose that the trouble with materialism is that it omits a representation of some element of the object of which it is a picture. The Cartesian ego, viewed as ectoplasmic, neither has nor could have a place in such a picture. The omission is perhaps one of detail, perhaps one of ignoring certain perspectives, or of supposing that no other picture is there to be drawn. Adopting an
incomplete picture of Humankind can lead one to become an incomplete human. Thus my motivation for this study appeals to such notions as the richness of human life; this will not in general be true of my arguments, though I would not regret it if it were.

If it be asked what positive views I hold on the nature of mind, I can only say that I hesitate to align myself with any of those philosophical camps which purvey theories of mind. If I have a view which amounts to more than a series of commonplaces, it is not envisaged in either of the purportedly exhaustive schemes of possible positions in the philosophy of mind constructed by C.D. Broad [1] and D.M. Armstrong [2]. I shall, however, not shrink from giving some account of mental concepts and it will become increasingly apparent that this owes much to the later thought of Ludwig Wittgenstein [3] [4] [5] [6] [7].

Having laid a few of my cards on the table, I should make a few remarks about the history of materialism and about how this is related to the main topic of the essay. It seems to me that the history of the concept of matter can be characterised as an interplay between two distinct senses of the terms in whose use the concept is embodied. If we take "materia" in "materialism" in an Aristotelian sense, we might take the view as one which stresses the sense (if there is
one) in which things can be said to be identical with their constituents, or one which supposes that if things consist in other things, then it is only the latter which can truly ("really") be said to exist. These positions have their analogues in the various reductionist strands of contemporary scientific realism. As materialism exists today, however, the notion of matter which is important is more closely akin to that of what Aristotle calls "prime matter", the "ultimate" constituent(s) of things. For Aristotle, of course, "matter" and "constituent" are relative terms; they refer to different things depending on what is the substance of which those are the constituents, and on what are the conditions of its coming to be and passing away. The idea of matter as a kind of stuff, of which everything (or at least everything material) is ultimately composed, though hinted at in Aristotle, is derived from Pre-Socratic philosophers such as Empedocles and, perhaps more notably, Thales. It is this second idea whose survival is more prominent in the notions of matter appealed to in Descartes and Locke in the eighteenth century. Thus I would locate the philosophical origins of modern materialism in the patchwork notion of "material substance", itself the mutant progeny of earlier concepts, which has emerged from the mutual misunderstandings of the British Empiricists.  

Through the success of physics and chemistry we do now
have a fairly respectable notion of matter (or mass/energy) and of the (relatively) ultimate constituents of material things. Problems arise when this notion is taken over by philosophers who, apparently unaware of the importance in its history of what I claim are philosophical mistakes, try to talk about it in ways which are best suited to the original concept of matter as this occurs in Aristotle.  

The gradual restriction of the application of "matter" to the "ultimate" constituents of physical objects has resulted in the near-synonymy of "materialism" and "physicalism" within present philosophical discourse. So the general cluster of views at whose downfall I aim could roughly be summed up via the catchphrase "Everything is reducible to the objects and laws of physics". It is not necessary to the view that we restrict ourselves to the theoretical entities and the relations between them which are at the fore of physics at the moment. The view is open-ended in so far as it admits of progress in physics. However, if the view is to have any content as a thesis of ontological parsimony, it must place some restriction upon what can count as an entity of physics, and the implicit restriction seems to be to the entities of micro-analysis. To a certain extent we can ignore these subtleties in doing philosophy of mind, since it is empirically likely that the laws which govern the
neurophysiological determinants of bodily movements can be stated adequately without reference to anything smaller than (or theoretically less primitive than) atoms and molecules. Without doubt most of the materialist contenders-in-the-field envisage this kind of success for neurophysiology.

The slogan by which I earlier characterised materialism/physicalism is multiply ambiguous. Its most important interpretations in the philosophy of mind are situated along a continuum whose poles we may identify as the contingent identity theory and eliminative materialism. The former view seems to be losing in popularity on account of some apparently insuperable objections. The latter, while set up to avoid those objections, has never gained much popularity; it is nevertheless a serious contender, since it has many theoretical advantages over the "Australian Fallacy", if I may presume to call this by an old name. My task in this essay is to argue against a certain version of eliminative materialism, the version outlined by Richard Rorty in "Mind-Body Identity, Privacy and the Categories" [10] and "Incorrigibility as the Mark of the Mental" [11], and defended by him in "In Defense of Eliminative Materialism" [12].

I focus on this version because, even though Rorty himself has since abandoned it, I regard it as the philosophical progenitor of other possible versions of
eliminative materialism; much of the work done in more recent years on eliminative materialism (e.g. that of S.F. Savitt and R.K. Shope, which is discussed in Chapter 5) is consistent with, and can be seen as an elaboration of, Rorty's view, even though the philosophers who produced it would not take themselves to be committed to every detail in Rorty's seminal papers. Furthermore, in so far as the view appeals to the alleged incompatibility of neural and mental discourse (see Chapter 2, section (b)), it has the advantage over more "Feyerabendian" versions of grounding this incompatibility in what I later call the conceptual commitments of the two forms of explanation, and not in the alleged incommensurability of the relevant "theories". Rorty's view has a similar advantage, I would argue, over Paul Churchland's [51] position, which appears to offer no reason, over and above the superiority of neural theory, for the elimination of mental states.

A final reason for selecting Rorty's version as object of attack is that it allows a consideration of the confusion of the ontological status of the mind with its epistemological role. Dealing with this confusion, which appears to characterise traditional Cartesian accounts of the mental (and to some degree, as I argue later, the present version of eliminative materialism), allows me to formulate by contrast a
more plausible positive non-materialist view.

Rorty's view may be summarised as the conjunction of the following theses:

I. Mental terms are theoretical terms, as explicated in II-IV.

II. The primary function of mental terms is in the causal explanation of behaviour.

III. It is possible that, even given what we know about people's behaviour and the circumstances in which they find themselves, mental states do not exist.

IV. Alongside "mind-theory" there are competing, and incompatible, theories of the aetiology of behaviour, such as neurophysiology and functionalist (machine-theoretic) accounts.

V. Given the theoretical superiority of one of these competing theories it would be rational to deny the existence of mental states.

VI. V is true because of the theoretical incompatibility
(mentioned in IV) which rules out the identity theory. This incompatibility is the direct result of an incompatibility between the predicates "mental" and "physical".  

VII. The incompatibility between the predicates "mental" and "physical" is grounded in the incorrigibility of certain mental reports.  

In Chapter 2 I develop this picture of Rorty's view through an extended consideration of his "demon" analogy. This requires me to offer a brief characterisation of the core of the identity theory, in opposition to which Rorty developed his version of eliminative materialism. Chapter 3 begins the argument against this view by a consideration of the way Rorty employs the notion of incorrigibility in its support. In Chapters 4-6 I move to a more general set of arguments against the conception of mental terms as theoretical, these arguments involving uses of mental terms which are not causal-explanatory. The course of my argumentation hence flows roughly upwards through the list of theses by which I have characterised the view I attack. A more detailed summary of the argumentational strategy can be found at the end of Chapter 2.  

One preliminary point of interpretation must be dealt
with before I can proceed, and this is the matter of the understanding of the word "mental" in the above expression of Rorty's view. Is the target of elimination, in his view, mental states conceived of in a common-sense non-philosophical way, or the states of some (Cartesian) non-physical substance, with their attendant philosophical baggage? On my reading of Rorty, it is the first target which is important.

My reasons for this are several. In the first place, the second construal of Rorty, as opposing only mental states as conceived of as states of a Cartesian substance, leaves him in agreement with myself and most contemporary philosophers of mind on all the crucial ontological questions. In particular it is impossible to distinguish between his version of eliminative materialism and the contingent identity theory: both have the same ontological goals. It is true that Rorty sometimes minimizes the difference between these two views, but when he does this he is questioning the importance of worrying about whether to identify mental and neural states or to eliminate the former in favour of the latter. He is sometimes inclined to say that these two conceptual moves come to the same thing. Whether or not this questioning is consistent with his other stated views, it would be unfair to construe "mental" as "non-physical" here, because then Rorty would be wondering whether his eliminative programme were the
same as one which is clearly conceptually incoherent: non-physical states cannot be identified with physical states.

Secondly, Rorty [11] argues at length that the traditional features of non-physical substance, such as non-spatiality, cannot be attributed to mental states, with the exception of incorrigibility. It is his explicit view that the distinction between physical and mental is grounded only on that between corrigible and incorrigible, and not on any of the traditional distinctions associated with the Cartesian model of the mind. He attempts, furthermore, to give an account of incorrigibility which avoids any implication between being the proper subject of an incorrigible report and being mental "in the Cartesian sense." He does say in a footnote that the common concept of the mind is "irredeemably Cartesian", but it is not obvious how this should be cashed out. I myself have in the past been inclined to attribute to Rorty the view that the use of mental language commits one to the existence of Cartesian substances and their states, but I have been persuaded to abandon this interpretation by S.F. Savitt, who has pointed out to me that Rorty does not use (nor need to use) this view in his argument for the incompatibility of mental and physical. Even if he held it, it is irrelevant to the issue at hand, and this is my third reason for resisting the interpretations according to which
the target of elimination is, for Rorty, mental states construed as states of a Cartesian substance.

Accordingly my own use of "mental" in this essay must never be taken as having those Cartesian implications. When I talk of the difference between mental and physical I mean the commonsense distinction between having indignation and having indigestion, between being overbearing and being overweight, between feeling and falling. I am convinced that Rorty's use of the term is the same, and that he is arguing that it could be rational to deny that mental states, in this sense, exist.

A result of the restriction of my focus to Rorty's view is that my thesis takes on somewhat of an historical cast, and I wish to address this point before proceeding. While Rorty's version of eliminative materialism was developed by him until the second half of the last decade, and by such writers as Shope [35] until its end, I can find nothing in the literature of the 1980's which falls squarely into the mould formed by what is distinctive of Rorty's view, namely the insistence that elimination is more plausible than identification because of an incompatibility between mental and neural talk, an incompatibility which is based on incorrigibility. I wish to confront, then, the possibility that I have backed a loser, as it were, in selecting which view to attack.

If this is a fault, I can do little to remedy it within
the framework of the present essay. Against the above possibility it might be said that although in the last eight or so years scholarly discourse has turned away from Rorty's version, it is perhaps too early to declare it inferior to the other versions which presently enjoy popularity. In fact I believe it to be superior, for the reasons explained above (pp. 5-7); I also believe that it is a more interesting view, and one the consideration of which leads me along more significant conceptual paths.

The comparison of Rorty's view with Churchland's [51] is instructive here. As I understand his position, Churchland, in arguing for the potential replacement of mental states, relies on the pure intuition that theoretical superiority leaves this as the most rational option. Those who share the intuition would of course agree, but many do not share the intuition. Rorty attempts to justify that intuition on conceptual grounds. His view may be easier to attack, in so far as it attempts more, but it is not clear to me that this is a fault. On the contrary, it adds to the challenge and the significance of considering the view.

There has been a handful of articles critical of Rorty's position written during the 1980's. While some of these echo my conclusions in this essay, the grounds adduced in them are different from those which I rely on. Thus
Rosenthal [120] argues against the claimed incompatibility of mental and physical in favour of the plausibility of an identity theory. Richardson [119] and Double [117] have both argued against the possibility of an eliminative view like Rorty's dealing with the intentionality of mental states. I have not raised this issue because Rorty himself makes no attempt to deal with it, asking his readers to suspend inquiry into this aspect. In fact I do not think the intentionality of mental states is an insuperable bar to eliminative materialism, though I cannot do justice to the issue here. (See Churchland [116] for some thoughts on this.) My project is focused essentially on the "epistemological" formulation in terms of incorrigibility.

This brings me to the final few articles I wish to mention. There was a dispute in the literature in the late 1970's about the precise interpretation of Rorty's view - see Bush [53], Lycan and Pappas [58] and Cam [115] - in particular about how strong an elimination was proposed. I argue in Chapter 2 for an interpretation of Rorty as proposing the possibility of the strong elimination of mental states as well as mental language. To me the textual evidence is so clear that I have not thought it worthwhile to go very deeply into that dispute. Hiley [118], however, has criticised the above three authors, and the tradition of interpretation of Rorty in
which I myself stand, on the grounds that Rorty intended his view not as a stage in the development of materialism, but merely as an epistemological attack on the notion of the Given and foundationalism. My position on this issue is clearly stated in my conclusions in Chapter 7: this is the grain of truth in eliminative materialism. However, any sensible reading of Rorty's key articles must include the understanding that his goal extends beyond the epistemological to the ontological: he is attempting to demonstrate the possibility of the non-existence of mental states.

To summarise this review of the more recent literature, I shall say that nothing has been written in this decade which furthers the support for Rorty's view. What has been written on the eliminative materialist side has been a development of the Churchlandian branch. While the significance of Rorty's view has not been overlooked by critics of eliminative materialism, their criticisms have not focused on the incorrigibility issue, nor have their grounds involved the considerations of non-constative and justificatory uses of mental language. Thus there is a considerable difference between their attempts and mine to refute Rorty's view. As a result, I have felt justified in not giving these writings a place in my treatment of the issue.
Footnotes to Chapter 1

1. See Aristotle [8] 1029A.


3. An example of this conflation is arguably to be found in U.T. Place [50], where he characterises the claimed identity between mental and neural as expressible by "the 'is' of composition". The intent here is "modern", but the example offered as analogy, that of the table which is an old packing-case, is Aristotelian.

4. Certain versions of physicalism, e.g. that of Hellman and Thompson, are non-reductive. I discuss this option in Chapter 7, section (b)(iii).

5. While there are differences between the views expressed in each of these three papers, these are less significant than the common thread which runs through them and which justifies my talking of "Rorty's view". The major differences are treated in the discussion of incorrigibility in Chapter 5.


7. I say this because, on the surface at least, the incommensurability of two theories seems to imply their logical compatibility, not incompatibility. This can be correct despite the truth of the claim that incommensurability inhibits joint viability, in some more pragmatic sense. The issue is a complex one, to follow which would lead us too far afield.

8. Theses I-IV I attribute to Rorty on the basis of his appeal to Sellars' Myth of Jones and of his use of the demon analogy in explicating the role of mental language. While, in discussing the former, Rorty speaks of the original role of mental terms being theoretical, it is clear from the context that the only difference between the original and present roles is the addition of the reporting role which may fall into desuetude as the result of neurophysiological inquiry. For Rorty the explanatory role seems to remain the more important of the two. Particular passages which support the attributions are: I: Rorty [10] p.179,11.33-36; II: Rorty [10] p.179,11.36-38; I and II: Rorty [11] p.411,11.13-16,22-24,39-41; p.413,11.38-40; III: Rorty [10] p.181,11.3-5; p.182,11.16-


Chapter 2: Explication of Rorty's View

(a) Introduction

The view which I wish to consider in this essay, and to which I shall for economy of expression refer as Rorty's view, arose out of dissatisfaction with the contingent identity theory as an account of mental states.¹ My first task then is to offer a brief account of the more sophisticated versions of the latter position and to show how Rorty's version of eliminative materialism might seem attractive as an alternative. Through a careful consideration of Rorty's demon analogy, I develop an account of Rorty's view and of how the dispute between these two forms of materialism might be settled. This allows me to formulate the strategic plan of my own attack upon Rorty's view, which I pursue in the ensuing chapters.
(b) The Contingent Identity Theory

The form of the contingent identity theory which I shall consider contains two main parts: a theory of meaning for mental terms and an empirical hypothesis. J.J.C. Smart's original version of the meaning-theory was expressed in his appeal to topic-neutral translations: mental ascriptions were rendered in the form: "Something is going on which is like what goes on when . . . ", the lacuna being filled with a physical description of the environmental situation of the relevant sentient being. David Armstrong [2] fills the lacuna with a description of certain sorts of behaviour, and David Lewis [14] [15] incorporates the best of both worlds, by identifying mental states via their role as causal intermediaries between environmental input and behavioural output. Since the identification of mental states usually involves the consideration of both input and output, I shall regard Lewis' view as naturally superior to the others and as their representative in the discussion which follows.

The theory of meaning for mental terms then comes out as expressible with the use of the schema "Mental state M" = "Whatever has typical causes $C_1, \ldots, C_n$ and typical effects $E_1, \ldots, E_n"$, or with the use of a multitude of such schemata,
since which effects result will depend upon which causes occurred. This theory of meaning is really a theory of reference, and is more clearly expressed in the form "We use the term "mental state M" to refer to whatever has typical causes \( C_1, \ldots, C_n \) and typical effects \( E_1, \ldots, E_n \)." Articulated in this way, the theory is consonant with Sellars' [17] Myth about Jones, the man who invented the mind. The idea is that we recognise that some intermediary exists between elements in some series of causal chains, the intermediary usually being required to account for irregularities within some overall regularity. The mental terms which we now use refer to these intermediaries before we know what they really are, and we eventually discover (or will discover) that they are neural states. It is the last claim - the empirical hypothesis - which completes the contingent identity theory.

Rorty, as eliminative materialist, diverges at this point, for his belief is that whatever the intermediaries are they are not referred \( ^4 \) to by terms in present mental language. These terms refer \( ^4 \) to mythical theoretical entities which are about to bow out of the arena of psychology in favour of neural states. In one small way, the basic account of the meaning of mental terms remains nominally the same: one view regards mental states as successful theoretical constructs, the other as unsuccessful theoretical constructs. This is the
rationale behind Rorty's [10] characterisation of the two views as the "translation form" and "disappearance form" of the identity theory. It also explains his penchant for referring to theoretical replacement as one form of identity.\textsuperscript{5}

It should be noted, however, that the two forms of materialism here compared employ rather different conceptions of the theoretical nature of mental terms. For the identity-theorist these terms are topic-neutral: anything could be discovered to be the referent of a mental term, provided it stand in the correct causal relations. For Rorty, however, mental terms are (and must be) in a sense theory-laden: it is part of the concept of a mental state that it have features incompatible with its being a physical state. This difference is crucial to an understanding of the superiority of Rorty's view over the traditional forms of the identity theory.

As Rorty ([11] p.401, fns.7&8) points out, these views have typically foundered on the existence of too many counter-examples to the schematic topic-neutral "translations" of mental ascriptions which they propose. He thinks he has an argument which demonstrates the superiority of his eliminative alternative over the various options he outlines as possible reactions to these objections. Again, each of the latter strategies gives an account of the meaning of mental terms according to which it is an open question whether the
referents of those terms turn out to be physical items, perturbations in some ghostly aether or what-have-you. The objection to this is that allegedly it leaves us with no workable distinction between "mental" and "physical", no range of referents across which to stretch this bipolarity, and therefore with no sense to either of the polar terms. These neo-Kantian claims drive Rorty to the conclusion that there must be some way of specifying the difference between mental and physical by supplying a property which all mental things have (even if in some cases only derivatively) but which all physical things do not. Thus he gives an account of mental concepts according to which anything mental, and nothing physical, is the proper subject of an incorrigible report. This is the major part of Rorty's claim that common sense talk about the mind is irredeemably Cartesian.

Similar objections to the identity theory have been raised by philosophers, such as Kurt Baier [18], who are not materialists. They raise deep issues which will receive treatment later in this essay; for the moment I will attempt simply to defuse Rorty's version of the argument. The first problem is that it is not clear that "mental" and "physical" are polar terms, each of which gets its sense from a contrast with the other. Rorty ([11] p.402, first paragraph) supports the claim by identifying the physical with the material, the
mental with the immaterial, and noting that "material" and "immaterial" are such polar terms. Even allowing that, we can question the identification of the mental with the immaterial, for economies, constitutions, even propositions, are immaterial without thereby being mental. Thus even if the structure of the argument is sound (which I doubt), this particular application of it will not work.

Even though I reject Rorty's argument to the conclusion that topic-neutral accounts of the mental must be flawed, I am prepared to admit that his view is superior to them in so far as it is not subject to the same objections by counter-example, for no attempt is made to offer "translations" of mental ascriptions. The same thing can be said in relation to a further set of objections brought against certain forms of the identity theory by, among others, Norman Malcolm [19] and Tyler Burge [20]. These objections, in my opinion the most convincing attempts at refutation of the cruder versions of the identity theory, turn on the idea that the truth-conditions for the ascriptions of many mental states include, on obvious inspection, facts about the world other than the central nervous system of the being to whom they are ascribed, and that therefore these states (or the fact that they exist) cannot be identical to states of, processes in (or the fact that there exists a certain state of, or process in) the
central nervous system.

I shall not here repeat the arguments of the philosophers who have raised this objection. Both this and the objections previously mentioned are directed more at the empirical hypothesis incorporated into earlier versions of the identity theory than at the topic-neutrality theory of meaning for mental terms that is the major conceptual component. An identity theory which does not limit the scope of the possible referents of mental terms to the central nervous system of the relevant subject can evade the second type of objection quite easily, it should be clear. Similarly, there are strategies available to the identity-theorist for overcoming objections of the first type, as Rorty himself notes. If I am right in claiming that Rorty's argument from the alleged polar opposition of "mental" and "physical" is unsound, then the move to some sort of eliminative materialism does not appear to be forced. In my discussion of the demon analogy I shall enquire whether this provides us with an independent argument for the superiority of Rorty's view to any form of the contingent identity theory.

Of course, even if there were no argument for the superiority of Rorty's view, it would be worth elaborating as an hypothesis which differs in important ways from the identity theory. Distinctive of eliminative materialism, in
the form in which it was originally espoused by Rorty, are some claims about the truth-conditions for the ascription of mental states which are written into the use of mental language. Rorty [11] thinks that it is importantly characteristic of mental concepts that certain reports in which they are employed are incorrigible. This is the key feature which, for him, allows us to separate mental and physical. Granted this conception the argument goes through very quickly: at some time in the future there will be no such incorrigible reports; therefore there will be no mental states. Here the first premiss represents the results of a successful neurophysiology, one which finds no need to appeal to the Cartesian picture of the mind to yield a complete account of the causation of human behaviour.

I shall postpone until Chapters 4 and 5 my criticisms of the general theory of meaning for mental terms which treats them as theoretical terms. The initial purpose of the present section is to examine in greater detail the grounds which might incline one to favour Rorty's version of eliminative materialism as a solution to the mind/body problem. I shall consider Rorty's [10] model of theoretical change - the demon analogy - noting that positions analogous to that mentioned, as well as to other positions in the philosophy of mind, are possible with respect to the demon/germ dispute. I shall
argue that one can decide between these competing positions on the basis of a detailed analysis of the use of language both prior to and posterior to the introduction of a new theory. A fortiori the dispute concerning Rorty's version of eliminative materialism can be decided on the same basis. Here the eliminative adjunct to the theoreticalist theory of meaning for mental terms, that certain mental reports are ("necessarily"?) incorrigible, is crucial. After pointing out some difficulties in the assessment of the plausibility of this claim, I postpone that task until later chapters.
In the hope that the reader will sustain the above picture of my strategy through what follows, I turn now to a consideration of the demon analogy (see Rorty [10]). The basic story concerns a primitive tribe the members of which "explain" disease by reference to different-coloured demons. The demons can be "seen" under various conditions (usually intoxication resulting from the ingestion of sacred mushrooms) by the witch-doctors of the tribe. We are to assume, though it is difficult to see how, that there is a reasonable amount of consensus on such beliefs as what colour of demon causes what disease and on what form of exorcism will best effect a cure; without this assumption it is a distortion to attribute a theory of disease to the tribe. Into this conceptual setting comes the civilised anthropologist and her medical team, who oppose the demon-theory, bringing evidence in favour of their own germ-theory. If the team is more successful in explaining, predicting, controlling and curing disease, and if in addition they can supply an explanation of the so-called sightings of the various demons, then the demon-theory will gradually be replaced by the germ-theory. Those members of the tribe who become educated in the new ways will decide that
there never were demons, and that whatever methods were formerly employed in the treatment of disease, the results achieved were coincidental, accidental or explicable in terms of the new theory.

It seems eminently reasonable, given the situation as described, to deny the existence of demons, and so it is with sensations and other mental states, according to Rorty. The first step in questioning the closeness of the analogy is to discover whether in the demon case there could be some analogue of the identity theory which could be a plausible theoretical alternative. This I now do.

Let it be assumed from the outset that the germ-theory is correct, just as in the mental case it is assumed that we will at some time in the future be confronted with a correct neurophysiological theory of behaviour.

Now suppose that some bright son of a witch-doctor, wishing to preserve his parent's prestige, argues as follows: "Your new theory, which speaks of germs in the body of the patient, is good and effective but does not show that the core of our old theory was false. On the contrary, it shows that it is true, for the two theories are identical. We use different terms for the principal causal agents in disease - you 'germ', I 'demon' - but in fact demons and germs are identical. All you have really shown us is that demons are
very small."

This suggestion is an intriguing one which can be developed in any of several different directions. What these are can be discovered by attempting to argue against it. One way to go - the "conceptual commitment" route - is to claim that demons have, by virtue of the nature of the concept "demon", properties incompatible with their being germs. Apart from the problem with size, demons are supposed to be personal in some sense, coloured depending on the disease, able to move around freely in this space and others, and so on. It seems that none of these properties pertains to germs, and so (by Leibniz's Law!) germs and demons cannot be identical.

Here our tribal identity-theorist can make several moves, all of which have their analogues in the debate in the philosophy of mind. He can avoid the objections concerning irreducibly supernatural properties by recasting his proposed identities, in Nagelian [16] fashion, in terms of a person's body being inhabited by germs and its being inhabited by demons, for none of the quoted properties differentiates these states-of-affairs. Still, if he wants to show that this is more than sleight-of-tongue, he must supply an account of the original use of the demon-theory which makes it seem natural to regard this as a primitive version of the germ-theory.
What this amounts to in this case is isolating some common beliefs about demons - that they are personal etc. - from any connection with the concept of "demon". This move is analogous to denying of the concept of mind that it involves incorrigibility.

My insistence upon the centrality of the issue of what is part of the concept of "demon" might be seen as slippery philosophical practice, for we have become accustomed, in the post-Quinean era, to regarding it as impossible to distinguish between those truths about the reference-class of some term which are true conceptually (or in virtue of the meaning of the term) and those which are merely well-supported empirical generalisations. For the time being I shall leave aside the cluster of arguments gathered by Quine and others in favour of this conclusion; I merely state my opinion that the arguments do not show as much as is desired. My justification for relying on the "belief/concept" distinction is that, in the case under discussion, it seems to me that the plausibility of Rorty's claims about what goes on when demon-thought is rejected in favour of germ-thought depends upon the assumption that certain truths are part of the concept of demon. What I mean by this is that, for most speakers of English at least, something which was not personal, for example, would not count as a demon, and this for reasons which have nothing to do with
the weighing of evidence for or against any empirical hypothesis about demons. The reader is clearly intended by Rorty to read such implications about demons into the tribal use of "demon" and this is where the difficulties arise, for there is a question, given the theory of the witch-doctor's son, about whether the tribe uses "demon" in the same way as most other English speakers.

Difficulties, but not impossibilities. What is at issue is whether "demon", as this is used by most English speakers, is an adequate translation of the tribal homophone, the primitive identity-theorist proffering "whatever causes disease" as an alternative. It does not seem to me that it is always (or even often) impossible to decide between two such alternatives, though the project of setting out to do so requires some basis for communication between the members of the tribe and members of the expedition. Let us assume that this exists, whatever account we wish to give of its genesis. Then the obvious way out of our difficulty is for the team to discuss the matter with members of the tribe. "Is there (could there be) such a thing as a demon which was not a person, or not very like a person in certain respects?" If the tribe was unable to handle the modal version of this question, the team would have to ask how the tribe came to know that all demons are personal, if that is what they believe. Here the
answer might be along the lines of "Because that's what we mean by 'demons'" or might be along the lines of "Because demons cause disease and anything which causes harm to a person in this manner is wicked and only persons and things like them can be wicked". There are of course many other answers, but the issue between the team, as eliminativists (who wish to reject demons on the grounds, partly, that they are "essentially" personal), and the tribal identity-theorist could be settled by seeing which of these two answers was the closer approximation. (If neither, then we need an alternative account of the meaning of "demon".) The fact that when the witch-doctors "see" a demon around the body of a sick person what they "see" has more or less the shape of a human being is important, but can be accommodated by either theory of meaning.

I have been somewhat optimistic about the success of these investigations, and I have to admit that it is possible that no unequivocal answer will be reached. There may, for example, be little or no consensus about how their term for demon is used amongst members of the tribe. The point I wish to emphasise is that where there is consensus about use, coupled with agreement in use, the anthropological linguist has a handle to grip in her attempt to decide whether the tribal identity-theorist is correct in his "topic-neutral"
account of the meaning of "demon". It seems obvious that facts about the prior use of language and the reaction to the germ theory will be of immense importance in helping her to decide.

It is difficult to give a precise account of what detailed conditions would have to be fulfilled for the demon-germ identity theory to be correct. The general picture, however, would be something like this: there must be a certain kind of continuity in the use of "demon" and related expressions before and after the discovery of the putative identity. Most of the statements about demons which were considered true before the change will also be considered true after it, with or without the substitution of "germ" for "demon". There will of course be some statements which change in truth-value with the introduction of the new theory, but if these are few enough in number, or perhaps theoretically insignificant to a sufficient degree, it will be relatively easy for the tribal doctors to amend them without too much strain on the demon-germ identity. On the other hand, if the number of beliefs which must be dropped is large, or includes the most important beliefs about demons, this is evidence that with the introduction of the new theory came a new theoretical term, "germ", which cannot be regarded as synonymous, nor even co-extensive with "demon". 7
One necessary condition which it will be important to mention of the appropriateness of a demon-germ identity theory is the existence of (almost) exceptionless correlations between demon- and germ-ascriptions. I shall not pause here to inquire into the form such correlations must take (in the mind/body case this has been the subject of much debate). This condition on its own is not sufficient for identification, I would argue, but combined with some of the others mentioned above it may be.

Along with linguistic continuity, and partly constitutive of it, one would expect some continuity in medical practice. This implies that in their choice of remedies, the witch-doctors had already anticipated, to a fair degree, the treatments which are sanctioned and recommended by the germ-theory. The anticipation may be more or less of a coincidence, but a radical break in therapy would seem to betoken a radical break in theory. If, for example, the previous treatment for pneumonia was chanting "Ngudul" while waving around the head of the patient sticks with small bells attached to the end, whereas the present one is the administering of antibiotics, this indicates that the doctors conceive very differently of the (two?) projects of getting rid of demons and getting rid of germs; and this is some evidence towards the conclusion that for them demons and germs
are not identical. To settle the question one would need to know more about their justification for the original treatment (since by the lights of modern science it can hardly be justified by success!) but it seems fairly clear that any such justification which could save the demon-germ identity theory would involve pretty bizarre beliefs (about the causal relations between words and microbes), themselves likely to be inconsistent with the germ-theory.

(In the context of the issue of successful explanation and treatment of disease, I wish to add a short digression on the suitability of the demon-analogy for Rorty's general case. Even were this issue not complicated by the possibility, not treated by Rorty, of an analogue to the contingent identity theory, it is not obvious that the analogy is close enough, in the crucial area of explanation and prediction, to be useful. For from our present scientific perspective it seems that the demon-theory should have no predictive value whatever. If this is so, then it is unclear what would have to be true for us to be justified in treating demon-talk among members of the tribe as the expression of an explanatory theory of disease, as opposed, say, to irrelevant religious rigamarole. Assuming that this problem can be overcome - if, for instance, demon-talk is sufficiently similar and similarly-treated to other areas of tribal discourse which are indisputably theoretico-
explanatory - I see no reason to resist the conclusion that
the germ-theory should rationally be accepted on the grounds
of its success in providing treatments and cures for disease.
Unless the witch-doctor takes this step s/he is left with the
fact, glaringly difficult to explain, that the germ-theory
functions so well, even if in his/her view merely as an
heuristic device. For similar reasons s/he may be obliged to
accept the hallucination explanation of the so-called
observations of demons (though this is to accept something
other then the germ-theory).

On the assumption that the demon-theory is
explanatorily useless, it is thus easily rejected, but then
the demon analogy is not close enough to the case of the mind
for it to prove illuminating. Mental entities, by contrast
with supernatural ones, do appear to be successfully used in
something which at least passes for the explanation and
prediction of behaviour, even if this something does not yield
many predictions which are incontrovertible. If we suppose
that this feature of mental talk is in fact shared by demon-
talk, it is not so easy to dispose of the latter in the
absence of an explanation for its partial predictive success.
Such an explanation is not provided by Rorty in the demon-
case, but must be forthcoming in the case of the mind lest we
be left in the uncomfortable position of the witch-doctor.
It is now time to assess the relevance of my extension of the demon-example to the case of mind-brain identity. The moral of the discussion is that whether or not the introduction of a novel theory (with novel theoretical terms) is a replacement or an embellishment can be decided by detailed reference to the facts, linguistic and other, which constitute the use of the relevant expressions both prior to and posterior to the purported theoretical change. Furthermore, these same facts are what are relevant to an assessment of how Rorty's version of eliminative materialism itself stands vis-a-vis the mind/body problem. This is perhaps hidden from us by the fact that the demon-analogy, while it provides for the possibility of an identity theory as an alternative to elimination, is not intricate enough to allow for the formulation of any plausible analogues to dualism, phenomenalism or the view which would be analogous to my own, that the demon-theory and germ-theory are compatible in a way which does not involve the identity of germs and demons. The demon-example looks more like an open-and-shut case of theoretical superiority; the mind/body case requires a wider range of consideration. Since the example is so schematic, I have been able only to hint at the important
kinds of consideration, but I hope to be able to say more in relation to the case which is my real object of interest.

My general conclusion about Rorty's appeal to the demon-analogy is that while the analogy is not close enough to the mind/body case to provide strong support for Rorty's view, it is rich enough to allow for an elaboration - the one I have made - which indicates the sorts of considerations relevant to the acceptance or rejection of that view. I pursue these in what follows.
(d) The Conceptual Commitments of Mental Language

As soon as we turn from the fabricated world of example to the real world of the mind/body problem we encounter a number of further difficulties. Prominent amongst these is the obvious fact that there is little or no consensus about how mental language is actually used, what conceptual commitments it involves and so forth. (That is the question.) This is not surprising when one considers the fact that we are dealing here with a vast and variegated area of language and not with the use of a single word, as in the demon case. It may be reasonable to agree with Rorty that the common conception of the mind is Cartesian, where this means that most people think of statements about the mind as involving a commitment to incorrigibility, but it is less clear how significant this is to the question of whether the concepts of the mind and of the mental involve a commitment to incorrigibility. I am here assuming a distinction between the concept of mind and conceptions of the mind. The former is that the grasp of which is manifested in the use of mental language in the normal, correct way (including with certain commitments). The latter is a set of beliefs one would assent to, beliefs about minds and mental states, about mental language and about mental concepts. For every individual,
there is an issue about whether her or his conceptions are or are not in line with the concept enshrined, in part, in the commitments of the use of mental language. The complexity of the issue seems to be too great for the ordinary user of the language to cope with in a systematic manner, unless s/he be trained in some such discipline as philosophy or psychology. As a result there is a greater gap than in the demon case between what commitments are involved in the use of mental language and what commitments are believed to be involved, so that the latter is only tenuous evidence for the former. Therefore Rorty's claim that the concept of mind involves a commitment to incorrigibility is only tenuously supported by the fact that many or most people so conceive of the mind and its states.

I make the further supposition that conceptual commitments arise out of the use of language only in so far as language is a communal or intersubjective activity. Thus the beliefs of the particular user about the referents of her/his terms do not on their own determine what commitments are involved in her/his use of them. To the extent that we share a language we share a conceptual scheme. Commitments arise where people agree in the use of language (where their usages coalesce) rather than where they have the same beliefs about the semantics of their language, though the latter place may
be on the evidential route to the former.\textsuperscript{10}

I will not argue for this supposition, nor the theory of language in which it is embedded, at this point; I hope to support it by the plausibility of the results it yields in application to the problematic at hand. If the supposition is correct, then, we can ignore the lack of consensus about what it is to use mental language, as a theoretical physicist can ignore the lack of consensus about what it is to use quantum-mechanical language,\textsuperscript{11} and continue our investigation into what commitments are involved in that use, conceived as an intersubjective enterprise.

In any case, despite the difficulties involved in specifying the conceptual commitments of the use of mental language, it seems clear that Rorty must appeal to some such notion in the defence of his view. Since at the time of writing "Mind-Body Identity, Privacy and the Categories" Rorty did not acknowledge this point,\textsuperscript{12} it would be well for me to examine the grounds he mentions there for preferring elimination to identity. These are, in the case of Xs being identified with/replaced by Ys (Rorty [10] p.184),:

\begin{enumerate}
\item The Y-laws must be better at explaining the kinds of phenomena explained by the X-laws (not just equally good). Indeed, they must be sufficiently better so that the inconvenience of changing one's linguistic habits by ceasing to make inferential reports about Xs is less than the
inconvenience of going through the routine of translating one's X-reports into Y-reports in order to get satisfactory explanations of the phenomena in question.

(B) Either Y-reports may themselves be made non-inferentially, or X-reports may be treated as reports of mental entities. For we must be able to have some answer to the question "What am I reporting when I non-inferentially report about an X?", and the only answers available are "you're reporting on a Y" or "you're reporting on some merely mental entity."

Rorty proposes these only as necessary conditions of the advisability of elimination over identification; as such they are fairly plausible. However, an argument for elimination requires sufficient conditions, as Rorty appears to acknowledge in writing "Incorrigibility as the Mark of the Mental", and these cannot be supplied by A and B. The main reason for this is, I suggest, that they ignore the possibility that confronted by the superiority of Y-laws to X-laws we may take the option of preserving Xs by changing the X-laws to redress the imbalance. This is merely a special case of the survival of an entity through the restructuring of beliefs about it. (See, for example, Putnam [23] for an application of this notion to the concept of acid.)

This discussion may appear to be vitiated by the fact that Rorty admits that (A) will not be fulfilled in the case of sensations, but I do not think that this is so. For this
admission is inconsistent with Rorty's general view, since it involves retreating to a position (which Rorty calls "traditional materialism") which does not distinguish between the identity theory and eliminative materialism. If all that is important to materialism is that "at no greater cost than an inconvenient linguistic reform, we could drop [mental] terms," then materialism is trivial, for any term could be dropped, it might be argued, with no greater cost than inconvenience. Furthermore, it does not follow from a term's being dropped that its referent does not exist; so if we are to take Rorty seriously as an eliminative materialist, this formulation of the view he is expressing is much too weak. And Rorty does argue for the "disappearance theory".

A second point to notice about (A) is that strictly it will not be fulfilled if perfect correlations are established between mental states and brain states. As Rorty says ([11] p.423):

If such correlations occurred, every explanation in terms of mental states would be isomorphic to an explanation in terms of neural states, neither mode of explanation being simpler or more elegant or more fruitful than the other.

It would seem that under these conditions it would not be convenient to eliminate mental states as opposed to
identifying them with neural states. However, there seems to be an inconsistency in Rorty at this point, for immediately following the above passage he insists that the existence of correlations is a necessary condition for the elimination of mental states (via the overthrowing of incorrigibility). This echoes a footnote in the previous paper (Rorty [10] p.177,fn.) which makes the same claim. The appearance of inconsistency is heightened by the concluding remarks in the latter paper, where Rorty claims that, given a parallelism of the form described above, "[i]nsistence on the "identity" of the mental and physical would seem to be an unnecessary rhetorical flourish." For if the flourish is merely rhetorical, then it is irrelevant whether it be made or not; but this is to say that there is no significant difference between parallelism and identity, in which case parallelism can hardly be a necessary condition of the eliminative alternative to identity.

It seems to me that the best reaction for Rorty would be to drop the reliance of his view on psychophysical correlations, and this is the position which I shall consider him to have taken. Apart from being a way out of the inconsistency outlined above, this move has other advantages. It enables Rorty to sidestep any objections, such as those of Malcolm and Burge mentioned in section (b), which stress the
overwhelming empirical likelihood of its being impossible for neural science to come up with true correlations of the relevant kind. This is a further way in which the thus-amended eliminative view of Rorty is superior to traditional identity-theories, though of course the failure of parallelism is not sufficient for its justification.

I conclude that Rorty's initial discussion of the demon analogy does not provide him with sufficient grounds for justifying his version of eliminative materialism. If elimination is to be preferred to identification this must be because of the conceptual commitments involved in the entity-to-be-replaced. In the demon case, it seems to me that the reason for eliminating demons is their possession of properties (size, personhood) incompatible with their being germs. Whether the term "demon" survives is a matter of the convenience of the wholesale restructuring of beliefs; whether if it is retained it will refer to the same things seems more of a foregone conclusion: it will not - instead it will refer to germs. Whatever the merits of the case I make here for the elimination of demons, it seems to be what Rorty requires in the mind/body case, and what he commits himself to in "Incorrigibility as the Mark of the Mental".

It is important to note, of course, that even if Rorty were correct in his argument against the identity theory, this
by itself would not establish his view. It seems to me that it is possible that the predicates "mental" and "physical" be incompatible in the way Rorty desires, so that nothing falls under both predicates, and that there exist a complete physical (neurophysiological) explanation of behaviour, yet that mental states exist. A useful comparison here is of the mind/body case with that of the relation between structural and functional (hardware and software) explanations of the behaviour of complex machines. In an unpublished paper [24] I argue that what we have here is the existence of two parallel causal explanations which are not incompatible, despite the fact that it may not be plausible to identify functional with structural states. (I do not argue this point here.) At the very least it seems plausible to suggest that functional terms would have application whether or not identification were possible, and therefore we are entitled to say that functional states exist. So too do mental states, in my view.

It is not important for me to argue for this view in the present context; all that is necessary is that it be a possible view, for then Rorty will need further argument in order to justify his claim that the success of neurophysiology as an explanation of behaviour renders it rational to deny the existence of mental states. Thus I am inclined to suppose that Rorty is committed to something like Thesis IV, whether
or not he grounds the theoretical incompatibility of ordinary mental talk and neurophysiology in the alleged incompatibility between mental and physical, as I have supposed. Rorty gives us a story about how we might stop talking about mental states, but it is not at all clear how this would impugn their existence unless Thesis IV were true.

Rorty himself does not acknowledge that the parallel explanations account is possible. The only alternative he offers to elimination or identification is, in the demon case, the possibility of attributing disease to "the compresence of demons and germs", and he points out, probably correctly, that this tactic falls to Ockham's Razor. But this is not the alternative I have suggested, as should be clear. I shall return to the assessment of Thesis IV in Chapter 4.
(e) Strategy of the Attack

In Chapter 3 I tackle Rorty's use of the notion of incorrigibility in support of his view. I shall distinguish between the "disappearance" and "non-existence" forms of this view, and explain why I give prominence to the latter. I shall argue that (i) mental ascriptions (first-person, present-tense etc.) are not incorrigible in Rorty's sense, nor in the traditional sense; (ii) Rorty's form of incorrigibility fails to distinguish between mental and physical, in that it applies to some physical ascriptions; (iii) even if Rorty's form of incorrigibility were the mark of the mental, it could not ground the categorial incompatibility between mental and physical which he seems to require; therefore Thesis VII is false.

Having treated the argument in favour of Rorty's view, I move on to a consideration of to what extent the view accounts for the phenomena of mental discourse, assessing it now in terms of its explanatory value. Rorty himself distinguishes two roles of mental language, the explanatory and the reporting uses, the replacement of which is a necessary part of the elimination of the mental. I suggest that this is a partial view.
In Chapter 4 I argue that reason-explanations are distinct from, though related to, causal explanations and that they pose a problem for the interpretation of mental concepts as theoretical, this notion being first explained. A discussion of the significance of this problem for Rorty's view is postponed until Chapter 6, except that here I adduce some reasons for doubting Thesis IV.

In Chapter 5 I discuss the reporting role of mental language, arguing that (i) Rorty's supposition that mental reports can be construed as reports about brain-states is false; (ii) in so far as mental reports are incorrigible, this cannot be accounted for on Rorty's (Sellarsian) view; (iii) Rorty has failed to supply adequate support for Thesis VI (and hence for Thesis V); (iv) this removes a large part of the justification for believing Thesis III.

In Chapter 6 I look at non-constative uses of mental language, and the existential commitments they involve, as well as the causal commitments involved in reason-explanations (as described in Chapter 4). I argue that (i) as a result of the existence of these uses of mental language Thesis II is questionable, if not false, and so is Thesis I; (ii) even if we allow Thesis II, it is by no means clear that the uses described above can be accommodated by neural talk (in particular, the adoption of neural talk will necessitate
significant changes in the structure of our systems of moral justification); (iii) hence, whether or not we should use mental language is not merely a matter of theoretical superiority coupled with linguistic convenience; (iv) it would be rational to maintain the use of mental language for the functions described above; (v) these uses commit one to the existence of mental states (though this is fairly trivial given the falsity of Thesis VI); (vi) therefore, Thesis V is false.

In Chapter 7 I shall summarise the argument and draw my final conclusions about Rorty's view. I shall be satisfied that I have dealt adequately with its main thrust if I can establish that Thesis V is indeed false. I pass on now to the first step in this plan.
Footnotes to Chapter 2

1. Throughout this essay I use the terms "mental states" and "brain-states" without regard to the distinction between states, processes, events, even properties, except where one or more of these distinctions becomes important in the consideration of some particular view. Thus a creature is in some mental state whenever some mental predicate is true of it - I leave open for now the question of what qualifies as a mental predicate.

2. Lewis' view is actually more sophisticated than this since some of the laws which bind mental states involve causal connections between mental states. Smart's and Armstrong's considered views seem to differ from his only in emphasis.

3. The only improvement of major significance which could usefully be made to Lewis' version is the incorporation of Thomas Nagel's [16] suggestion that psychophysical identities should be considered to hold not between states but between facts. I do not believe that this amendment would greatly alter the substance of my comments on the contingent identity theory.

4. The situation is actually more complicated than this. See the discussion of different senses of "refer" in Chapter 5.

5. As Rorty [11] puts it:

   ...the sense of identity in question is the sense in which [caloric] is identical with (is replaced by, is eliminated in favour of) the kinetic motion of molecules. (p.401)

(I insert "caloric" in place of "phlogiston" because the latter notion, though it had a role in former explanations of combustion, was only arguably relevant to the matter of heat content.)

   Although it seems to me that there is no such sense of "identity", that is mostly a verbal matter: it does seem to be a result of modern theory of heat that caloric (and phlogiston) do not exist, and I find it odd that someone should wish to identify the kinetic motion of molecules with something which does not exist, at least in the absence of some strong justification for the analogy thereby suggested between replacement and strict Leibnizian identity. These
thoughts, however, leave the real issues untouched.

6. This formulation is ambiguous. The claim might be that when mental reports cease to be regarded as immune to challenge mental states will cease to exist, or that when this happens they will be discovered never to have existed. Throughout this essay I give prominence to the second of these alternatives, for reasons which will be explained in the following chapter. Rorty [11] himself thinks there is nothing to choose between them, but I argue later that this is incorrect.

7. Here I appeal to some ideas (about meaning and conceptual change) which I cull from Wittgenstein. See for example Wittgenstein [5] ## 79 and 87, though these passages are subject to other interpretations.

8. At this point it may be worth while to inquire into the feasibility of a witch-doctor's rejecting the germ-theory and clinging instead to the traditional view. This stance would correspond, perhaps, to dualism or phenomenalism in the philosophy of mind. Rorty's supposition is that it would be irrational to reject the new theory if it provided better prediction and control of disease, along with an explanation of the observations which are taken to support the old theory. The latter is accomplished not indeed by the new theory itself but by another part of the more general medical viewpoint in which it is embedded, in this case psychopharmacology, since the accusation is made that the alleged sightings of demons are merely the products of hallucination. Of course, such an accusation is hardly likely to convince any witch-doctor who has his/her wits about him/her, since s/he can parry it with the claim that the scientists hallucinate through their microscopes, or if the concept of hallucination is not available to him/her, at least s/he can claim demonic possession! Ultimately, I suppose, the coherence of the scientists' possessed rantings will settle the issue in their favour.

9. The direction hinted at by Hooker [21] as possible for a theory of retention/replacement seems not to be applicable to either of the present cases, as, on the surface at least, it requires that the old and new theories be formulable in the mathematical sciences. (See pp.223-4.) Perhaps some analogue of his suggestion could be constructed for the mind/body case - I cannot tell yet whether this would be consonant with or in conflict with the view I am here endorsing and attributing to Rorty.
10. These very general remarks must be qualified, for they do not apply at the borders of our conceptual field where the creative development of our concepts is often in the hands of a few or even of one; however, mental concepts have been sufficiently incorporated for this subtlety to be ignored.

11. I have in mind the opposition between the Copenhagen and other interpretations of quantum mechanics.

12. That Rorty did eventually come to rely on this argument is evidenced by (i) his use of the notion of incorrigibility in "Incorrigibility as the Mark of the Mental" - see my footnote 9 to Theses VI and VII in Chapter 1 - and (ii) his admission, in a letter to Richard Sikora - see Sikora [22] p.192,fn. -, that this is the case.

13. I have some quarrel with (B). The second alternative, that X-reports may be treated as reports of mental entities, has no application to scientific advances outside the demon analogy and confuses hallucinations with their contents. Reports of demons are not reports on hallucinations - they are reports on nothing. This alternative, in the case of mental states, also could offer no explanation of second- and third-person reports of mental states. Clearly the first alternative is the only one plausible, and a discussion of Rorty's appeal to the idea of non-inferential reporting on brain-states will be found in Chapter 5.

In relation to (A) I should mention that while I should resist the notion that generalisations about mental states are laws in a scientifically rigorous sense, I shall allow this mode of expression to Rorty for the sake of argument.
Chapter 3: Incorrigibility and Incompatibility

(a) Disappearance versus Non-existence

As I noted in Chapter 2, section (b), Rorty ([11] p.421-2) thinks there are two possible scenarios for the overthrowing of the incorrigibility of certain mental reports which would justify his version of eliminative materialism. Before I examine his arguments in favour of the relevance of incorrigibility to the mind/body problem, I wish to explain why, finding one of these scenarios much more plausible than the other, I give it prominence in the ensuing discussion.

The two possibilities are (i) that when cerebroscopes come to be used to override sincere first-person present-tense mental reports, the states thus reported on will, losing their incorrigible status, cease to be mental, and (ii) that when mental reports cease to be incorrigible, it will have been discovered that there are no mental states (nor ever were). The second alternative is the one outlined in "Mind-Body Identity, Privacy and the Categories" and is, I wish to argue, much more plausible than the first.

My reasons for this are as follows: first, even though on the first alternative mental states would lose their status as mental, they would continue to exist, that is to say,
sensations, thoughts, beliefs, emotions and so on would exist though they would not merit the epithet "mental". To an opponent such as myself of the elimination of mental states the loss of this word seems a small price to pay for the continued existence of those states!\(^1\) To put the point less dramatically, it is hard to see how this scenario constitutes the elimination of mental states from the ontology of the universe, as opposed merely to their redescriptiion. To regard the first alternative as viable is to regard eliminative materialism as an epistemological thesis (concerned merely with the overthrow of incorrigibility) as opposed to an ontological one, and this seems incorrect.

Second, if mentality is a temporary status, it is hard to see how this status constitutes the categorial gap which Rorty seems to require between mental and physical. Suppose that with the advent of cerebroscopes people learn to self-ascribe neural states, but that they never need to be corrected, or at least that the practice comes into being of never questioning these ascriptions. According to Rorty's claim that such incorrigibility is the mark of the mental, it would seem that under these conditions neural states will have become mental; yet it seems undeniable that they are physical states also.\(^2\) This is not however allowed by Rorty's general views about the opposition between mental and physical, and
for the sake of consistency he must reject the first alternative in favour of the second.

This, then, is my justification for regarding the second alternative which Rorty offers as the representative of his views in the discussion which follows. Where possible, I shall try to indicate whether my arguments against his views articulated in accordance with this option are relevant to the rejected interpretation, even though this appears inconsistent with Rorty's more general views. I begin by arguing that mental reports of the relevant kind are not incorrigible in the sense outlined by Rorty, nor in the traditional "logical" sense.
(b) Rorty's Form of Incorrigibility

The sense in which Rorty [11] believes certain mental reports to be incorrigible, and this incorrigibility to be the mark of the mental out of which arises the fundamental gulf between the mental and the physical, is a reinterpretation of the original notion. Basically, he shies away from the idea of logical incorrigibility (see Rorty [11] pp.414-5) which has historically been associated with experiential reports, and for good reason: for if experiential reports are thus incorrigible, any sincere ones will be true, and it is hard to see how Rorty could reconcile this with his belief that possibly there are no mental items, experiences etc.. Instead he uses a sense of "incorrigible" in which some set of reports is incorrigible if there are no accepted procedures, in some linguistic community, at some time, for overthrowing them. The difficulty here is that this definition appears to let in as incorrigible much that is not mental, for the phrase "no accepted procedures" is somewhat vague. In various different senses of it, there are, for example, no accepted procedures for overthrowing a person's report that s/he spoke with God yesterday, that s/he has indigestion or is constipated, but we would not say that these reports were of
mental events. In a society with a total taboo on nakedness, there might be no accepted procedures for disputing persons' descriptions of their own bodies. If it is replied that in such cases there always could be such procedures, it might be worth noting that this sounds like a relapse towards the notion of logical incorrigibility. Rorty owes us an account of the limits of acceptability, in the sense which he thinks is here relevant, before we can judge whether or not his definition is too broad. I believe that it would not be possible for any such account to deal adequately with the theological example, in any case, and so am led to conclude that indeed it is too broad.

On the other side of the coin, R.I. Sikora [22] has pointed out that there are accepted procedures for overthowing people's claims to be perceiving after-images with a certain number of sides. It appears that Rorty's definition of incorrigibility perishes between the Scylla of excessive narrowness and the Charybdis of excessive breadth.

In response to the above criticism by Sikora, Rorty did amend his notion of incorrigibility [25], but at a price, as Sikora [26] has argued in rebuttal. According to Sikora's original counter-example we would overrule someone's claim to be having a fourteen-sided (say) after-image if we knew that the physical object which caused the after-image had (say)
only thirteen sides. Rorty admits this but insists that if
the subject persisted in the claim that the after-image had
fourteen sides, we should eventually be forced to accept the
claim and to declare a fault in her/his visual "apparatus" or
in the theory explaining the occurrence of after-images. This
"repeated report" form of incorrigibility is the one which
Rorty, in the end, considers to be distinctive of the mental.
As Sikora points out [26], however, the scenario which is here
supposed to persuade us that mental reports of the relevant
kind are incorrigible in some sense is precisely that which,
according to Rorty [11], would be required to overthrow the
idea that there exist states which can be reported on
incorrigibly, except that the decision to favour report over
theory is arbitrarily reversed. Rorty's attempt to make the
incorrigibility criterion of the mental plausible eats away
directly at the plausibility of his claim that mental reports
will cease to be incorrigible at some stage in the development
of neural technology.

It seems to me, then, that this exchange between Rorty
and Sikora, as well as the considerations I mention at the
beginning of this section, together refute the claim that the
incorrigibility (in either of Rorty's senses) of certain
mental reports is distinctive of them and them alone, and
therefore that Rorty has no support for his further claim that
incorrigibility, being the ground of the alleged incompatibility between mental and physical, can be used to justify eliminating mental states. Furthermore, even if there were no counter-examples to the application of Rorty's notion of incorrigibility, his argument would still not be successful. The problem is that the existence of the "empirical" form of incorrigibility is too weak a basis upon which to found the incompatibility of mental and physical: from the mere fact that some reports are not questioned it does not follow that they cannot be about physical entities. (If, for example, it became the practice not to question remarks made by Prime Ministers, this would not mean, alas, that a report of cruise missile testing in Alberta made by Brian Mulroney would be a report about merely mental events.) Perhaps this is tacitly admitted by Rorty at those points at which he is inclined to express his view (paradoxically) as the view that mental states will cease to exist when mental reports cease to be (treated as) incorrigible. I discussed this option in section (a) of the present chapter.

I conclude that Rorty's way of arguing for Thesis VII does not and could not work. There is no logical connection between incorrigibility (as he means it) and the incompatibility of mental and physical. Hence, as Rorty intends Thesis VII to be taken, it is false. In Chapter 5 I
shall make some further remarks about the relationship between what I call the theoreticalist aspect of Rorty's view and his treatment of incorrigibility in an attempt to show that the latter cannot be encompassed within a theoreticalist framework. For now I am content to refute his actual argument.
(c) Traditional Forms of Incorrigibility

By way of a digression, I wish to add a few comments on the traditional account of incorrigibility. According to this notion, as found, say, in Descartes, it follows from A's being in pain that s/he knows s/he is in pain, and therefore that if s/he says s/he is not in pain (in some language which s/he understands), then s/he is lying, or at least attempting to deceive: it is logically inconceivable that A should be in pain and not know it. I do not think that this claim is viable, for if we had strong evidence of a neurophysiological kind for A's being in pain, and a well-supported neural theory of pain, we could overrule A's obviously sincere report that s/he is not in pain. It is important to note that this objection does not rest on any form of the identity theory; it merely presupposes some well-established correlations between people's being in pain and their brains' being in a certain state.

The "logical" form of incorrigibility is also open to Sikora's "after-image" objection, though this is, as intended, more limited in scope than mine above. While it is more plausible to suppose that the existence of this form of incorrigibility would betoken some kind of deep
incompatibility between mental and physical, I have already noted that Rorty cannot appeal to this notion, for if there are in this sense incorrigible reports of mental states, it follows that there are mental states, and this is inconsistent with Thesis III. Traditional forms of incorrigibility are then of limited relevance to the present problematic.
(d) The Criterial Form of Incorrigibility

Before I go on with the argument against Rorty's view I wish to record the fact that I do believe that certain mental reports are incorrigible in a certain sense, though this may not be as significant as was originally thought. I prefer to express the claim that incorrigibility is a feature of first-person present-tense mental reports (the incorrigibility thesis) in Wittgensteinian fashion; it then refers to the following mesh of criterial relationships:

(i) saying that one is in pain is a criterion of being in pain;
(ii) saying that one is not in pain is a criterion of not being in pain;
(iii) there are certain non-verbal behavioural criteria for being in pain;
(iv) there are certain non-verbal behavioural criteria for not being in pain;
(v) the joint fulfillment of the criteria mentioned in (ii) and (iii) is a criterion for one's not being able to speak the language of pain, or for one's being insincere in one's report, or for feigning;
(vi) there are independent criteria for linguistic ability, sincerity and feigning;
(vii) if the criteria mentioned in (ii), (iii) and (vi) are fulfilled jointly, we should conclude that one is not in pain, unless there are independent reasons for not doing so and there is an alternative explanation for the behaviour mentioned in (iii) (again governed by independent criteria).

This is of course only a fragment of the list of relevant criterial relationships (which is in any case open-ended). The list is consistent because the satisfaction of a criterion does not entail the existence of that for which it is a criterion. The possibility envisaged in (vii) of other criteria for being in pain is just the possibility of the relevance of neurophysiological facts and of their employment as criteria. There is nothing in Wittgenstein to deny (and much to support) the possibility that criteria may change to incorporate such new information.

The form of incorrigibility which I have outlined is midway between "logical" and "empirical" incorrigibility as I have described them, for it allows that first-person present-tense reports about pains may be overruled (at a later stage of scientific development) without supposing that it is a mere fact of sociology that such reports are not at present
overruled. It is a fact which is central to the application of the concept-cluster: "pain", "sincerity", "linguistic competence" etc..

This account stands in need of further explanation. While this may appear to involve a further digression, it is essential to at least one of my arguments in Chapter 5 that I offer some characterisation of an acceptable version of the incorrigibility-thesis, and it seems best to do that here for the sake of contrast with the other forms of incorrigibility mentioned. The first step in this explanation is to offer a summary of what I call the criterial view of mental language. This is taken from an unpublished paper by myself [27], though many of the basic notions are adumbrated in my M.A. Thesis [28]. I should emphasise the fact that, as it is presented here, this summary contains no argument for the criterial view.

The criterial view is a view about the meaning of certain terms, that is to say about their use: I have nothing to say against the existence of meanings, but as entities these are none too popular these days and I shall have less trouble if I characterise them in terms of use. I have in mind the conventional use of terms, their potentiality for acting as vehicles for human intercourse. The correct description of the use of criterially-governed terms, mental terms included,
is enormously complicated.

I begin with the rejection of two forms of account. There is some connection between the ascription of mental states and the attribution to the subject of pieces of behaviour, dispositions to behave and circumstantial position. What is it? According to one view rejected, there are entailment relations between mental ascriptions and circumstantial/behavioural attributions: this is (logical) behaviourism. It is rejected on account of a plethora of standard counter-examples, the most fatal being perhaps that supplied by Putnam [29]. According to the other view rejected, Rorty's version of theoreticalism, mental states are posited as causal intermediaries between circumstances and behaviour. Here the connection between behaviour and mental states is evidential, it being a clear possibility that something else could have taken on the job.

According to the criterial view the relevant connection is that of a reason to that for which it is a reason. The fulfillment of a criterion is a reason for ascribing a mental state. It is important to see that (and how) this claim is part of a theory of the meaning of mental terms. The theory should be expressed as follows: "Mental terms, at least in so far as they occur in mental ascriptions, are used in this way: their application to a particular case is taken to be
justified if a sufficient number of the relevant criterial statements are (or can be) known to be true." What justification amounts to must be explained in a non-question-begging way: a statement is justified if it is (or would be) considered (by most members of the relevant linguistic community) to be acceptable, pending any future change in the criteria for the use of the statement. There is thus a certain provisionality in the use of criterially-governed expressions.

The word "considered" here must be interpreted lightly, for I do not wish to say that criteria are established by people agreeing in their opinions about whether something counts as a reason or justification for making a certain attribution. They agree in using the criterially-governed expression in a certain way; they justify its application in particular cases by appeal to certain other facts. This is a coincidence in practice, not in opinion.

The fine detail of the account of the form of the criteria for mental terms is handled in terms of the concept of "pattern". Mental ascriptions are justified when certain patterns exist in the circumstances and behaviour of the subject. Consider the cluster of events which happen and conditions which hold when someone is in a mental state. That is a pattern. Connected with each type of mental ascription
is a disjunction of patterns, connections among which are established by bonds of similarity and family-resemblance; the disjunction is open-ended in so far as novel patterns may be added to it as they crop up (it is also possible for patterns to be cut loose). The open-endedness accounts in part for the failure of entailment relations between pattern and mental state. In learning to use mental language we learn to associate mental ascriptions with familiar patterns. Much of this learning is piecemeal.

Not all of a pattern is available to observation on any particular occasion. The neurophysiological part is hardly ever available and so is not used at present as a means of identifying familiar patterns. This may change. When only a small part of the circumstantial and behavioural parts of a pattern is available, we are inclined to regard this as merely evidence for the relevant mental state. Here judgment is less certain, for the same elements occur in many different patterns associated with many different mental states. An ascription based on such a small fraction is hence an hypothesis of a certain kind.

This kind of situation appears to add plausibility to the theoreticalist view of the meaning of mental terms, because we say we explain the behavioural evidence by the mental ascription, suppose that the mental state caused the
behaviour and so on. According to the pattern-view, however, this appearance is mere appearance. The form of explanation is not theoretico-causal: I explain the part of a pattern when I make a suggestion about what pattern it is a part of. The use of the causal construction is also of little significance: I may be allowed to say "His pain caused him to wince" but this does not have theoreticalist implications. (It should also be noted that not all of the relations between pattern-elements are causal: some may be merely those of concurrence and temporal contiguity.)

Of crucial importance in all this is the question of the possibility of an alternative explanation. When the observed portion of the pattern is small, alternative explanations abound, although many of them will be mental. When much more of the pattern is observable, this is not true - there are very few situations in real life where investigation of the pattern will not settle the question of the subject's mental state. (The exceptions could mostly be handled by neurophysiological inquiry, but the open-texture of mental concepts probably makes it impossible to settle absolutely every case except by linguistic decision.) Mental ascriptions hence admit of certainty.

This certainty is supported by an important theoretical consideration. One can explain why certain pattern-elements
are connected evidentially with a certain pattern - because they tend to occur together and it is linguistically economical to talk about common concurrences (familiar patterns). One can also explain why a fairly diverse set of patterns is associated with a single mental state - there are overlapping similarities between the patterns, particularly in the behavioural portion (thus supplying predictive power to mental ascription), and, perhaps most important, we, as human beings, wish to treat those who instantiate the various patterns in the same way. (Those who suffer from heartache deserve sympathy as well as those who suffer from headaches.) There is however no explanation (apart from the trivial linguistic observation) of why this collection of patterns is associated with this mental term. This is why the criterial view is a theory of meaning.7

I now turn to the form of incorrigibility which I wish to endorse via this view. The foregoing account of the meaning of mental terms deals exclusively with third-person ascriptions of mental states. These are assumed to be basic, in so far as the meanings of other mental statements are to be explicated in terms of the criteria for third-person ascriptions. (E.g., to say that sadness is an element in grief is to say that one who fulfills the criteria for the ascription of grief also fulfills those for the ascription of
sadness; to say that self-hatred resembles egotism is to say that the criteria for the two overlap to some extent.) The purpose of the following considerations is to account in criterial terms for the alleged incorrigibility of first-person present-tense mental ascriptions.

Some of these, of course, are governed by criteria in the normal way and pose no new problem - e.g. I might attribute to myself a subconscious hatred of Americans by observation of my behaviour in certain circumstances. The same goes for all so-called "dispositional" states. When I say that I am in pain, however, I do not base my judgment on the third-person criteria. Do I have a different criterion in my own case? On the view which I am supporting, the answer to this question is a simple "No": mental self-ascriptions of this type are based on no criteria at all. This is, of course, the view of Wittgenstein.

It is easily demonstrated that first-person ascriptions of so-called "occurrent" states are not based on criteria. On the one hand, to employ some state-of-affairs as a criterion, as a reason for justifying some judgment, I must make a judgment about the state-of-affairs itself (though this need not be a conscious process, nor one essentially linguistic). However, the judgment that I am in pain is based on no other judgment. On the other hand, the logical relationship
between a self-ascription and itself is one of entailment, which is enough to render it non-criterial according to the view of criteria here espoused.

The incorrigibility of certain mental self-ascriptions stems from their dual nature as reports and indicators. In so far as they are reports they can be true or false, accurate or inaccurate, known for certain by others, held to be questionable or denied. In so far as they are indicators they can be reliable or unreliable. A person's verbal behaviour, including his/her mental self-ascriptions, is part of the set of criteria for his/her being in a mental state (of the relevant kind), and since it is only part, it will not settle with absolute certainty the question of whether the person is in that state. What the incorrigibility-thesis demands is weaker than this unfulfilled requirement - it is that the self-ascription is to be regarded as reliable unless the self-ascriber is insincere in making the ascription or does not understand the language used in making the ascription. There are independent criteria for sincerity here, in the demeanour of the self-ascriber, his/her past trustworthiness etc., and so that qualification will be passed over without further comment. I should like to focus our attention on the criteria for linguistic competence.

The general claim to be urged (with some qualifications
to be introduced later) is that the criteria for linguistic competence in the making of the relevant mental self-ascriptions are such that to be counted as competent a person must ascribe mental states to her/himself, or be prepared to do so, in precisely those circumstances in which it would be correct (justified) for others to make the same judgment about her/him, i.e. where the relevant criteria are fulfilled. (This can be so even though the criteria do not enter into the justification of her/his own assertion.) All I have to offer in support of this claim is the observation that we only regard someone as competent, and, for example, cease training her/him, when this condition is fulfilled.

Now it must be remembered that the above paragraph is not an account of the necessary conditions for being linguistically competent in the making of mental self-ascriptions. Linguistic competence is a mental state and so ascriptions of linguistic competence are governed by criteria. Hence it is possible that the grounds upon which judgments of linguistic competence are based may change in the future, e.g. as a result of neurophysiological discovery. If that happens, there may be exceptions to the general claim urged above (that for one to be counted as competent one's self-ascriptions must accord with the ascriptions of others) and thus exceptions to the incorrigibility-thesis. It seems unlikely, however, that
things will change so much as to rule out the general incorrigibility of mental self-ascriptions.

The criterial account of incorrigibility has certain advantages over those previously considered. In the first place it represents incorrigibility as epistemologically useless, in so far as it is merely a criterial condition of linguistic competence, and hence provides for a host of ways in which one can question reports while admitting that they are incorrigible. One can therefore admit that incorrigibility is a general feature of mental self-ascriptions without being committed to the Cartesian/foundationalist programme in epistemology. In the second place, it offers an explanation of why there exists the social practice of regarding self-ascriptions as incorrigible, a question which is left unanswered by Rorty's eliminative materialist account. The answer is in terms of the utility of training people so that their self-ascriptions match up with the ascriptions of others (the latter being generally guaranteed by the fact that they rest on the criteria which figure in the account of the meaning of the relevant mental terms). Again it is not that we make a conscious choice to follow a cause of action which appears to us to have this utility; it's just that the development of mental language has turned out to be linguistically economical in this respect.
The third advantage of the present account is that it enables us to explain the restrictions upon the scope of incorrigibility which are required by the existence of counter-examples to the incorrigibility-thesis. One of these, already mentioned in section (b), is due to R.I. Sikora. The substance of the counter-example, it will be remembered, is that one's claim to have a fourteen-sided after-image, for instance, can be overruled by the observation that the physical image used to induce the after-image has only thirteen sides. Hence not all self-ascriptions of sensations are incorrigible. The criterial view has a story to tell about this situation: here we have a case where there is a gap between making a correct report and being linguistically competent (a gap which is missing in the case of pain-ascriptions, perhaps), for to make a numerical judgment of this sort certain, one must fulfill the criteria for having counted correctly, this being something more than understanding how to count. In this case it seems clear that the only applicable criterion of having counted correctly is the agreement of one's answer with the result of counting the sides of the original physical template. Hence we can see how the claim of incorrigibility should be denied in this example.

This, then, is the form of incorrigibility which I am prepared to admit exists. In Chapter 5 I shall return to this
topic in an attempt to see whether Rorty's general theoreticalism can account for the existence of incorrigibility (in any form). The argumentative, as opposed to adumbrative, purpose of the present chapter has been to show that the connection between incorrigibility and the alleged incompatibility of mental and physical is not as Rorty thinks. I think this has been achieved.
Footnotes to Chapter 3

1. Notice that in this context Rorty cannot appeal to the idea, to be discussed in Chapter 5, that mental reports are really reports on brain-states, for it is an important part of his view that those states cannot be mental, and so cannot lose their status as mental.

2. At first glance, at least, it looks as if the same could be said of any physical ascriptions which might cease to be questioned at some time or other, but I shall be charitable in restricting my objection to the general area of the explanation of behaviour.

3. I have heard Rorty in conversation appeal to a pragmatic theory of truth to support the claim that statements about (even reports of) sensations can be true even though there are no sensations. I do not know what he would say about the statement, "Sensations exist". I take up these matters again in Chapter 5.

4. I should be inclined to say the same thing about both of Rorty's forms ("single report" and "repeated report") of incorrigibility, although it is unclear to what extent he would agree with me, as noted in section (b). Except where he appeals to the latter form, he seems to agree with me that neural evidence could overthrow incorrigibility in any of the forms mentioned so far. We differ in so far as he does, while I do not, believe that such an overthrow would impugn the existence of mental states.

5. This view is taken from the mature work of Wittgenstein, though some of my interpretations may not be beyond dispute.


7. The lack of an explanation has further consequences. Most importantly, it places the general existence of mental states beyond dispute. Mental states are not to be identified with patterns, or parts of patterns, but nor are they something other than the patterns, for the existence of which the patterns are mere evidence. A mental state is that whose existence we are entitled to assert (pending conceptual revision in the light of future discovery) when a pattern of a certain kind exists. There is thus no possibility of a different and incompatible explanation, of a theoretical kind,
of the existence of a pattern. Mental states are here to stay. However, it is not relevant to argue this point here.

8. In normal circumstances. One can of course imagine very bizarre circumstances in which I might have to base my self-ascription of pain on (e.g.) neurophysiological evidence, but in ordinary life such circumstances do not occur.

9. Mental self-ascriptions are also expressions of one's mental state. This status may or may not be reducible to the other two.

10. My view is distinguished from the one Rorty [10] attributes to Wittgenstein by the fact that the condition here is criterial, not empirical. This point is developed in Chapter 5.
Chapter 4: Mental States as Theoretical Entities

(a) Introduction

In Chapter 2, section (b), I represented an important similarity between Rorty's version of eliminative materialism and the contingent identity theory as the claim that mental states are, in one sense or another, theoretical entities. This is the issue I wish to take up now. I shall speak indiscriminately of mental states as theoretical entities, of mental concepts as theoretical concepts and of mental terms as theoretical terms: whatever differences one can find between these formulations of the central topic are not of interest to me here. I shall argue that the claim is false.

It will be remembered that the difference between the two views in this matter was characterised by my saying that whereas the contingent identity theory regarded mental states as successful theoretical entities (to be identified with brain-states\(^1\)), Rorty's version of eliminative materialism regarded them as unsuccessful. The grounds for the latter charge were (i) the greater explanatory power of neurophysiology as an account of the causation of behaviour, and (ii) the incompatibility of neurophysiology with ordinary mental explanation. I have suggested that Rorty argues for
(ii) by claiming that the ordinary concept of mind is irredeemably Cartesian (in the sense involving incorrigibility), and the previous chapter was devoted to an attempt to counter this argument. My attack against (ii), if successful, provides some grounds for refusing to regard mental states as theoretical entities, in the sense espoused by Rorty, since such regard is all part of the picture of mental concepts competing with those of neurophysiology in the race to appear in a better causal theory of behaviour, but these grounds are slim.

That (i) is part of the putative reason for rejecting mental concepts raises a number of questions. The normal justification for the relevance of (i) is that the function of such concepts, and of the uses of language in which they are employed, is pretty well exhausted by the attempt to give a causal, theoretical account of human (and other?) behaviour. If mental terms had some further functions, it would seem that even in the face of the scientific success of neurophysiology we could preserve mental talk for those functions, unless indeed neurophysiological talk could come to fulfill those functions too. Ignoring this last matter for the time being, I take it, then, that the claim that mental states are theoretical entities implies the claim that the use of mental language is (pretty largely) exhausted by its role in the
causal explanation of behaviour. As I said before, I shall argue that the latter claim is false.

It is perhaps too naive to mention the obvious point that mental language did not grow up in the laboratory, but in the world of general human affairs. While it would be acceptable, even persuasive, to me to argue that the ontological and epistemological rules and conventions which govern the development of scientific theory are different from those which govern that of ordinary discourse, I do not expect points of this kind to carry much weight with proponents of the empiricist/scientific realist tradition. The institutional differences between theory-production and human life tend to be under-emphasised by the latter philosophers; although I think they are important, showing this to be the case is too long an endeavour for this essay, and I must proceed as if it could not be done. Perhaps the success of this work will help partially to dispel the illusion of the scientific homogeneity of human discourse.

The normal (and Rorty's) starting-point for discussion of the theoreticality of mental states is Sellars' Myth of Jones, the Inventor of the Mind [17]. We are encouraged to suppose, by this account, that the concept of the mind, together with other mental concepts, functions as if the following schematic history of its genesis were true: Caveman
Jones notices certain overall patterns of similarity in how other cavepeople behave in particular circumstances, and also exceptional circumstances where the patterns are altered, where the normal connections do not hold; in order to account causally for the exceptions, he proposes a system of intervening variables (beliefs, desires etc.); eventually the theory becomes popular, is extended and elaborated until it becomes the complicated structure of folk-psychology which is common to almost all human beings today. 2

Of course, since the above scenario is not supposed to be a true account of the genesis of mental concepts, one must ask what restriction the story is thought to place upon the present use of mental language, if we are to approach an appreciation of the point of the Myth. It would of course be possible for a term or set of terms to enter the language much as described above, but for it then to come to have a different role later. This possibility is not, I think, envisaged as part of the Myth. What the latter implies, then, is that the role of the concept of mind is pretty much now as it was in the story - a theoretical construct for the causal explanation of behaviour.

Let me mention some preliminary difficulties with this view. The first concerns the relationship between the mind and particular mental states. There are two possible general
lines which might be followed on this question. The first is that the mind is a theoretical construct at a deeper level than particular mental states - that is to say, first we explained behaviour by reference to particular mental states, then the latter by reference to a mind. This line I do not find intuitively satisfying: given a sufficiently complicated array of mental states attributable to an individual, I can see no logically possible alternative to his/her/its having a mind. What else could explain the presence of mental states? To suppose that the concept of mind is only one of a number of competitors in the theoretical battle for the causal explanation of the possession of mental states is to treat it in a Cartesian manner, as having some features other than being the underpinning for mental states (which might disallow, for instance, the identification of mental and physical). I have argued against this picture of the concept of mind. One of the great advantages of Sellars' Myth, one might add, is that its adoption enables us to escape the Cartesian picture of the mind: for instance, one avoids the conception that mental ascription is basically self-ascription in a way which makes other-ascription derivative and more-or-less doubtful. Reasons such as these suggest to me that it would be more consistent to reject this line of development of Sellars' view. I cannot however show that it is impossible to
take it.

The second alternative is to take X's having a mind as a logical consequence of her/his having mental states. Hilary Putnam [30] is one example of a philosopher from the empirical realist camp who finds this path attractive. It is insufficient to suppose that the ascription of any single mental state to an object implies the object's possession of a mind, for it appears to make perfect (even if a little metaphorical) sense to attribute a few mental states to computers which yet lack the sophistication to qualify as having minds. However, if the array of interconnected mental states attributable to an object is both large enough and complicated enough, I see no reason to deny the implication that the object has a mind. At the moment only human beings and higher animals, within our ken, qualify for this entitlement, but I see no reason why robots too should not achieve the standard at some later point in the development of cybernetics. On this view, "X has a mind" says nothing more than "X instantiates a large and complex array of mental states" - of course one can still argue about what mental states are.
(b) Actions or Movements?

The previous matter does not make difficulties for the view that mental states are theoretical entities. I introduce it only to show that the view requires clarification on one point and to foreshadow what I take to be a greater difficulty. This is the question of what exactly is to be explained by reference to mental states. So far I have been using the deliberately vague term "behaviour" to cover an opposition which must now be unearthed. "Behaviour" can be taken to mean "set of actions" or "set of physical movements" - which is relevant here?

The problem with choosing the first alternative is that a logical connection, analogous to the one illustrated above between the possession of mental states and the possession of a mind, might be argued to exist between being the agent of certain actions and the possession of mental states. The point is easily established if we do not require specificity in the account of what mental states are tied logically to certain actions, for then "having mental states" comes to much the same thing as "having a mind" and the possession of a mind is implied in the ascription to beings of such action-predicates as "... is going for a walk", despite the fact
that this is not a predicate describing the subject's mental states. Many ascriptions of physical actions are written in the language of the mental, the validity of the use of such ascriptions therefore presupposing the legitimacy of mental concepts. It is more difficult to demonstrate entailments between being the agent of an action and possessing some particular mental states, but it is also difficult to deny, for example, that most ascriptions of intentional action, while not referring to the relevant intention, bear some sort of implication that the action was undertaken with this certain intention. Think of such ascriptions as: "He pressed on towards the warmth of the cabin", "She murdered her sister", "He carried the child all the way to the hospital". It may be that some action-ascriptions do not have this character, but I am inclined to think that it is a general feature of the language of action.

The point I am making is not the Humean one, notorious in the context of the discussion "Reason versus Cause", that because of the logical connection between actions and mental states the latter cannot be regarded as causes of the former (I think this view is false), but the more modest claim that on account of this kind of connection, mental states cannot be taken to be theoretical entities for the causal explanation of actions. I take it that for Ys to be
theoretical entities appealed to in the causal explanation of some phenomenon A, it must be logically possible for A to occur even if there are no Ys (in other words, some competing theory is possible). I do not think that this condition is fulfilled in the present case.

Some ordinary-language descriptions of behaviour might be thought to fall between the categories of action and movement which I have separated. An example might be "Arthur smiled broadly". I do not mind calling smiling something other than an action, but I insist that the ascription of smiles shares the features which above I allotted to action-ascription. The sentence just quoted bears the implication, in its normal use, that Arthur is in some sense happy, pleased or joyful (or, arguably, pretending to be so). Thus it is not merely a description of the topography of Arthur's face. This can be seen by taking the two cases where Arthur is, first, a person and, second, a chimpanzee. When both of these Arthurs smile, the topographies of their faces have little in common, but their states of mind have much.

It seems to me that the only interpretation of "behaviour" consistent with the condition which I place on the theoreticality of causal explanations of behaviour is as "physical movements", where these are described in the non-question-begging language in which record-player arms can move
just as sensibly as human ones. I shall argue later that the systematic causal explanation of such movements is not primary among the functions of ordinary-language mental ascription and explanation. This interpretation does allow us to see how neurophysiology is a viable contender for the role of fulfilling this function, since it is conceptually continuous with the description of bodily movements.

I must now anticipate a general line of objection to these last considerations. Many materialists would, I think, argue that the dichotomy I have stressed - the opposition between actions and movements - is a bogus one, since it is possible in some way to reduce actions to movements. There are various accounts of what such a reduction would amount to, and thus various positions parallel to reductionist positions in the philosophy of mind. I shall consider two of them, analogous to the identity theory and a version of eliminative materialism, in the hope that this treatment will supply the general outline of an approach to all of them.

The "identity theory" would suppose, then, that actions are identical to physical movements, or, in a more sophisticated version, that the fact that an agent carried out a certain action is identical to the fact that her/his body moved in a certain way. Our attention here is restricted to those actions which involve (as I would put it) physical
movement; purely mental actions, such as calculating in one's head or deciding to learn German, can be ignored as of dubious status, except in so far as they can be reduced to the neurophysiology of the agent.

My objections to this version can be fairly easily stated. First, notice that type-type identities are implausible in this context, for a single action-type, for example "poisoning the inhabitants"\(^4\), can be accomplished through an enormously wide range of physical movements, and a single physical movement can be involved in a vast range of types of action - the movement which summons a friend to the door of her/his apartment is the same as that which might destroy the earth through the detonation of a nuclear device. The only way I can see to avoid this conclusion is to insist that the descriptions of bodily movements must include not only the facts about the motions of the agent's limbs, but also those about the manipulations of the physical world these involve. The problem with this respite is that the list of relevant manipulations may be endless, and it is intuitively implausible that such a list is part of the description of a physical movement, rather than of its effects.

Even if the alleged identities are expressed in token-token form, there are difficulties for this version of the "reductionist" view, for if my earlier argument is correct,
action-ascriptions have implications not only for the effects of the physical movements involved, but also for the types of mental state underlying the action, whereas descriptions of physical movements lack such implications. The arguments of Burge and Malcolm, mentioned in Chapter 2, section (b), show that, furthermore, mental states cannot be identified with neurophysiological states alone (nor mental facts with neurophysiological facts alone) because the implications which mental ascriptions have for the surroundings, social and otherwise, of the subject. So the general expression of my dissatisfaction with the "identity theory" version is that action-ascriptions have truth-conditions which spread out over a much larger portion of the spatio-temporal world than those of ascriptions of physical movements.

The second alternative, analogous to an eliminative materialism, would centre around the supposition that because of difficulties like the above, we should reject action-ascriptions en bloc in favour of physical-movement-ascriptions. This reduction-by-disappearance has the very clear-cut consequence that the theoretical role of mental language is in the causal explanation of physical movements. The relation of the latter to actions is, on this view, largely irrelevant.

This reductionist objection, then, while it may find
fault in my appeal to the concept of action in my earlier considerations, cannot force the rejection of my conclusion that when mental states are spoken of as theoretical entities, it is the causal explanation of movements which provides the arena in which such entities are put to the test. (The second alternative in fact requires this conclusion.) I shall now proceed to some more direct arguments against the sufficiency of this account of the function of mental language.
I remind the reader that the notion of "theoretical" relevant here has two aspects, the causal and the hypothetical. To say that an entity is theoretical is then to say that reference to the entity occurs in some system of causal explanation (Thesis II) and also that it is logically possible that, given what we know about other things, entities of this sort do not exist (Thesis III). The considerations I shall bring to bear are aimed at one or the other of these aspects, usually not at both.

I shall concentrate my attention first on the causality-aspect of the claim that mental states are theoretical entities. The thesis to be attacked, it will be remembered, is that mental states are posited as factors in the causal explanation of bodily movements, and that this is the best account of their role in human discourse.

There is such a thing as the ordinary-language explanation of behaviour; it is a varied and complicated assortment of activities which it is perhaps hopeless to contemplate characterising in a few broad strokes. It will be sufficient to my purposes if I succeed in showing what it is not - the causal explanation of bodily movement. Even this
claim must be qualified, for it would not surprise me if some
uses of mental terms had this sort of function; my target is
more the view that the primary and pervasive function of
mental language is causal explanation and hence that ordinary-
language explanation is primarily and pervasively causal
explanation. My reasons for thinking that this thesis is
false are expressed in the following sections.
(d) Actions and Movements (Again)

Beginning at the movement end of the explanatory chain, we can ask for an example in which some bodily movement is explained by reference to a mental state. I raise my arm and execute a number of quasi-circular movements with my index-finger. Somebody asks why: the reason is that I am feeling lonely. This explanation is undoubtedly incomplete - not even a devout causalitarian would question this - but what is important is what it would take to remove this sense of incompleteness. Let us supply some more relevant mental states - a desire to avoid loneliness, a belief that executing such a series of movements will enable me to fulfill that desire. We must note that the movements can only be expected to have that effect if they are undertaken in certain circumstances, namely the favourable position of a telephone, so for the full explanation we shall require the further belief that there is a telephone in such a position. These beliefs and the associated desires combine to form an intention and upon this I act (move my fingers).

If this explanation sounds strangely mechanistic, I must be forgiven, for that impression is the result of my trying to represent as causal explanation something which I
deny is such. The point I wish to make now is, however, more limited than this. The example illustrates the following feature of ordinary-language explanation of behaviour, that it accomplishes an explanation of my A-ing in terms of my beliefs about A-ing, my desires to A, my intention to A and so forth; the explanation fails, in general, unless what is to be explained is the object of the relevant mental states. What I find odd about the example above is that the relevant A is taken to be a very specific series of movements (this is what is to be explained) and it is most implausible to attribute to the agent intentions and beliefs whose object is that series of movements. The only reason for doing this would be to bolster some philosophical theory of action - nobody in ordinary life would attribute to me the intention to move my fingers according to some elaborated geometrical itinerary. In the relevant sense of "object" the object of an intention is an action (e.g. telephoning my friend), and this was omitted from the explanation.

Of course, that reference to an action is an essential stage in the explanation does not entail that the ultimate explanandum is not a physical movement, but it has certain consequences for the relationship between different stages in the explanation. So far these relationships, on the theory being attacked, have been referred to under the general
denomination "causal", but it is a mistake to hold that causal relations exist between all referents appealed to in the different stages. If reference to an action is an essential part of ordinary-language explanation of behaviour, and it is held that what is to be explained here is ultimately physical movement, then it sounds as if there should be causal relations between actions and movements, that is to say between an action and the movements in which it is carried out. But this is nonsense. However arguable it might be that actions and movements are identical, it is not in the least plausible that the former cause the latter.9

Alongside this observation, it should also be noted that we do speak of actions explaining movements. Sometimes when I ask "Why is she moving her hands like that?" the simple action-description "She is ringing the bell" may suffice as an explanation of her movement without its being true that her ringing the bell caused her thus to move her hands. The action-ascription is an interpretation of the movement, not a reference to its causal antecedents.

My calling this form of explanation "interpretation" is yet but to name a difficulty, not to resolve it. The point to be made is that there is in the area of action-theory a form of explanation which is not straightforwardly assimilable to causal explanation as we know it in the special sciences; it
is hoped that the observations which led to this conclusion will motivate the demand for a general account of action-explanation which can encompass both this form and that of the explanation in terms of mental states. I approach such a general account in the next section.

The results of the present argument are as follows: the form of eliminative materialism being dealt with represents the ordinary-language explanation of behaviour as causal explanation, as involving a folk-psychological theory which is (or will be) a direct competitor with neurophysiology in a causal/theoretical battle for the explanation of behaviour. It is an assumption of this view that, pace Feyerabend [32], the two theories compete to explain the same data, namely the physical movements of (human) bodies. However, it is essential to ordinary explanation in terms of mental states that actions appear in some stage of the explanation. The relation between actions and movements is explanatory, but not causal: that is to say, actions are not causal antecedents of the movements by which one carries them out. For this reason the explanatory route via actions cannot be regarded as homogeneously causal; therefore the ordinary-language explanation of behaviour is not simply causal. Having established some motivation for this belief I can turn to the positive side of my account of the ordinary-language
explanation of behaviour.
(e) Causes and Reasons

Up until now I have been speaking of the ordinary-language explanation of behaviour as if it were a homogeneous type. This pretence must now cease. If I am to justify my claim that ordinary-language explanation is not primarily and pervasively causal explanation, I must isolate from its central cases those which non-controversially involve causal explanation.

We do, in ordinary life, ascribe to both actions and bodily movements causes which consist in events and states-of-affairs in the world outside the agent or mover and which do not operate via his/her mental states. Alfred fell over because the 'bus swerved; Betty spilled the milk because the cow kicked her; Cyril burned the cakes because his car broke down on the way home. On the surface there is nothing excessively theoretical about such explanations; however, they are undeniably causal, arguably involve only events and states-of-affairs which have a physical description and, it seems, are easily subsumable under physical (or physico-chemical) theory - indeed of not too sophisticated a kind. Except in so far as my future remarks must allow for the existence of these forms of explanation, they will not be relevant to the subject at hand.
So the first restriction on the scope of "ordinary-language explanation of behaviour" is to explanations involving the mental states of the agent/mover. However, it must be noted again that much of such explanation is straightforwardly (and merely) causal in nature. There are cases where what is isolated as "the" cause is an event or state-of-affairs outside the agent yet where the causal route to her/his action passes through her/his mental states. David dropped the pan because it was hot (but he wouldn't have dropped it if the heat had not caused pain). In cases such as these we have no hesitation in supposing that the mental state causes the action/movement as well - that is to say, it is an essential part of the full causal story. For my purposes cases of this kind are not significantly different from cases where mental states are said to cause actions or movements without mention of any external causes for the mental states. Ellen turned red, screamed and put her fist through the wall because she was angry; Frank failed to notice the stop-sign because of his headache; Greta stopped and stared at the ground, because the thought had suddenly occurred to her that perhaps all was not well at the manor; Hal was led into error because he believed that God demanded the renunciation of human ties; as a result of the same belief, Ianthe lived a very lonely old age.
In cases such as these we have no difficulty in supposing that the explanation offered could in principle be complemented by one in terms of the neurophysiology of the agent and in terms of whatever other physical systems are relevant. Where I do see a difficulty is in the claim that the development of a successful neurophysiological (etc.) explanation would show that the ordinary ones instanced above were false. This strikes me as most implausible. My suggestion, once again, is that we have two parallel causal stories, each told in its own terms and each true, by analogy with the hardware/software case mentioned in Chapter 2, section (d). On the level of theoretical explanation, that which is explained by the physicalist story will be the embodiment in physical terms of the action or reaction which is explained in ordinary terms, but at a more informal level we can allow for mongrel explanations which mix the two modes.

It might be thought that my admission of a causal role for mental states in the preceding examples is at odds with my claim that mental explanation is not primarily and pervasively causal. Perhaps I need to modify my claim: the point is that there is an activity, called explanation of action, which does not consist in the supplying of a causal account of the action, but which involves mentioning the mental states of the agent (this is the giving of reasons for action), and that
this activity could ground our belief in the existence of mental states.

Some preliminary difficulties must be dealt with before I can give my account of this form of explanation. The early parts of the historical debate about the alleged opposition between reasons and causes were vitiated because the argument centred on the attempt (or the futility of the attempt) to find and ground a distinction between two different kinds of things, namely reasons and causes. It was thought by the proponents of the distinction that these kinds were mutually exclusive - a mistake which left them unable to cope with a multitude of counter-examples. The distinction I suggest is between two activities, two forms of explaining, namely giving a reason for an action and giving its cause, and hence between two functions for ordinary mental language. It is part of my view that the same mental state may occur in both forms of the explanation of some action: Justin took his umbrella because he believed it would rain, and Justin's belief that it would rain caused him to take his umbrella. However, some causes cannot be reasons, and vice versa. Ken's unhappy childhood caused him to become an alcoholic, but was no part of his reason for so doing. Lolita entered the convent to find happiness, but finding happiness did not cause her to act in this way.
My comments on this last example must be quickly qualified. When a reason for some action is expressed in terms of a goal, some future state-of-affairs which is sought but not yet actual, it is of course nonsense to take the state-of-affairs as part of the cause of the action. However we can attribute to the agent some desire, wish, hope or whatever for the attainment of the goal, and this will presumably have a role in the causation of the action. This observation suggests the general principle that it is a (conceptual) condition of some goal's being the reason for a particular action that the agent have some form of desire for the attainment of that goal and that this desire be (causally) effective in bringing about the action.

This principle about forward-looking reasons has an obverse about certain kinds of backward-looking reason, namely those which are described in terms of some fact about the world other than the mental states of the agent. Take the case where Michael fired Nicholas because he (the latter) did not turn up for practice. Here it is Nicholas' failure to attend which is given as the reason for Michael's action, and even though this failure is, in contradistinction to the case of goals, part of the cause of the action, we are entitled to demand the further condition that Michael be in some mental state which involves, at least, the belief that Nicholas did
not turn up for practice. Otherwise that could not have been the reason for Michael's firing him. In such a case it will be further true that the belief-involving mental state is causally related to the action.

The demand for a desire is more prominent in the forward-looking case and that for a belief in the backward-looking case, but in fact each case involves each demand. In the first, it is a further condition of the goal's being the reason for the action that the agent believe\textsuperscript{12} that the action will at least facilitate the attainment of the goal, and that this belief be effective in determining the action. Likewise Michael must desire to fire Nicholas and this desire must be part of the cause of the firing.

I am thus led to propose the following principle: externalised reasons (e.g. goals and facts about the external world) require internalised reasons, consisting of a belief/desire pair of which it can also be said that it is part of the cause of the relevant action. (This is what Davidson [33] calls a primary reason. With this part of his view I am in agreement. I differ in the account of the function of reason-explanations themselves: he believes, whereas I do not, that reason-explanations are causal explanations.)

It is the relationship between actions and internalised
reasons which is of interest to us in the discussion of how mental states are related to actions. Having allowed the necessity for a causal connection here, I now wish to claim that more is required adequately to characterise this relationship. I shall present (in dogmatic fashion) a brief theory of reasons culled from a consideration of the role of reason-giving in human interaction.
(f) Reasons and Interpretation

Reasons are given in response to questions, posed by others or by oneself; such a question asks why the relevant action was committed, or movement made. Expressed in the most general terms, the fact is that the questioner is missing the appreciation of some characteristic of the action or movement which the person questioned is asked to supply. (It is irrelevant whether either person is the agent.) Sometimes the seeking of this characteristic is motivated by the desire for knowledge of the causal antecedents of the action, and sometimes in these cases it is the mental states of the agent which are obscure to the questioner. At other times, the question is not merely a request for factual information of this kind, but also a form of challenge: what the questioner lacks is the appreciation of the action as rational, as the act of a rational agent.

What we call giving the reason for an action is a response to a challenge of this form. What the questioner gains as the result of a correct response to his question is an interpretation of the action or movement as rational. The forms of response, and hence the kinds of reason, are various, and dependent upon what kind of lack is suffered by the
questioner - there is thus a great deal of relativity to context and circumstances in whether some X is a (or the) reason for some Y. "Why did he raise his arm?" - "To start the race": here an action-description offers a rational interpretation of his movement; i.e. the movement is represented as a rational means to the end of starting the race. "Because he wanted to start the race" would in many circumstances supply the same interpretation by reference to the mental states of the agent. "Why did he start the race?" - "Because it was time"; "Because that's his job"; "Because he doesn't want to be fired"; "Because urgent business needed to be attended to". Responses such as these are appropriate to the request for a reason in particular circumstances, where the person questioned knows from the context of previous discourse what element is missing from the questioner's comprehension of the action. It does not need to be pointed out that the examples of reasons given here form a very mixed bag.

So too do the senses of "rationality", or the kinds of rationality, involved in the (requesting,) giving (and reporting) of reasons. At one end of the scale is the fairly simple "means-end" rationality of decision theory, where a person's action is judged to be rational to the extent that it is a successful means to the fulfillment of some desire of his
or hers. Here what counts as rational is relatively determinate - we are inclined to think that whether or not an action is rational in this sense is a factual matter; although there can be disputes about such a matter, these will be resolvable through the acquisition of new information. The same thing cannot be said, however, about the other end(s) of the scale, for example in the area of moral rationality. It does not seem to be a simple matter of fact whether it is rational to condone prostitution, and new information will not assist us in settling this issue. Moral questions such as these, and even more obviously broader ethical questions about how one should live, are open to rational disagreement - that is to say, there may be no single determinate answer to the question of which of two contrary positions is the rational one to hold. The claim that one of these positions is the only rational one to hold is then to be seen less as a factual claim than as an evaluative one.13

To return to the scale of senses of "rationality", we should note that not only are actions judged as rational in relation to the mental states of the agent, or in relation to a set of "externalised" values, but also the mental states themselves are judged to be rational or irrational. Sometimes the issue is the comparatively determinate one of logical consistency (e.g. among beliefs), sometimes the comparatively
indeterminate one of whether the possession of such a state (belief, desire, attitude, feeling, etc.) is rational tout court; this is clearly an evaluative matter in many cases. What all these cases, senses and types of rationality have in common is that they are tautologically involved with justification and to this I now turn.
(g) Rationality and Justification

To ask whether an action is rational is to ask whether it is rationally justified, but this can mean a number of things. Sometimes our interest is only in whether an agent has a reason for his/her action; we say the action is justified if it is in rational accordance with his/her beliefs and desires. This notion of accordance is normative—a proffered reason only counts both as a reason and as a justification of some action if we (in general) so regard it, would at least feel prima facie inclined to act that way given that reason— but it is not dangerously so. What I mean by this is that in these cases the constraints upon rationality do not to any great extent involve evaluative opinion. The desire to avoid hunger, coupled with the belief that eating potato crisps will alleviate hunger, is a reason (if anything is!) for eating them; the desire to avoid thirst, coupled with this belief, is not. In this preliminary sense of "rationally justified" I would further suppose that the desire to avoid thirst, coupled with the belief that potato crisps quench thirst, would be sufficient justification for eating them. Any attempt to criticise this thirsty agent upon grounds of irrationality would proceed from questioning the rationality
of his/her set of beliefs, in terms both of consistency and of what s/he ought, given his/her life, to have picked up about the world. It can be seen that further normative elements creep in here.

There are other occasions, notably but not exclusively in moral life, where our interest extends beyond the question of whether the agent had a reason for his/her action, or even a good reason for his/her action. We can admit that a person has a prima facie reason for acting in such-and-such a way, yet question whether her/his action is rationally justified. This may be because s/he has other reasons (other beliefs and desires) which indicate a different course of action and which are more important to her/him. Oliver (who had no thought of suicide) was irrational to rush into his burning house to try to save his daughter's teddy-bear, though his love for his daughter provided him with a good reason for trying to save the teddy-bear. (His action is consistent with the suggested relative weighting of his desires for life and for alleviating his daughter's sorrow if we suppose that the former was "shelved" in the heat of the moment.)

There are other cases where the question of justification arises for an action for which there is admittedly a reason, but where our concern is directed not at the fact that the agent has other reasons for doing something
else, but merely at the possibility that his reason does not justify his action. One type of case would be overreaction—the gamekeeper who shoots the poacher; another would be where we object to some institutionalised reason-action justification relation, such as the belief that capital punishment is warranted in cases of murder. People sometimes argue that jealousy is irrational—if it is, this is because the grounds for jealousy do not justify the reaction.

These "ultimate" questions of justification, not relative to the beliefs and desires of the agent in the simple manner of decision-theory-rationality, are evaluative questions, and it is by no means clear that they always have determinate answers. Perhaps this is a reason for thinking that a different sense of "rationality" is involved in them; perhaps not. What is important to my case is the characterisation which these cases provide of an enterprise, namely the justification (and its questioning) by (and in terms of) reasons, which is both prevalent and important in human society. Given that the mental states of the agent have the ability to justify her/his actions, we should conclude that mental language has here a function over and above the theoretico-causal explanation of behaviour. As will be seen later, the open-ended nature of this enterprise, the fact that in it not all questions have determinate answers, provides
some reason for thinking that it is resistant to the replacement of mental language by that of neurophysiology. These conclusions remain to be drawn in Chapter 6.

The immediate conclusions of the present chapter are that (i) in so far as mental language is employed in causal explanation, it need not be viewed as competing with neurophysiology (or any other proposed causal account), and (ii) given that the purpose of reason-explanation is justification, not causal explanation, there exists an important function of mental language the validity of which is not a scientific issue. These conclusions cast doubt on Theses IV and II respectively.

I turn next to the reporting role of mental ascriptions, in an attempt to discover whether this can be accounted for adequately by Rorty's view, involving as it does the claim that mental states are theoretical entities. Initially my focus will be on Thesis III.
Footnotes to Chapter 4

1. This loose mode of expression ignores the Nagelian reformulation of the identity theory in terms of fact-identity. This omission, undertaken on grounds of economy, is not of importance here. What I say could quite straightforwardly be reinterpreted as relevant to the claim that mental facts are theoretical facts.

2. This is both a simplification and an embellishment of Sellars' Myth, but neither qualification affects the use to which I put the Myth. The simplification is made because Sellars' version, where thoughts are to be postulated first as causal substitutes for spoken sentences, involves the questionable supposition that speech could be assigned a role in the causal explanation of behaviour in the absence of any beliefs about the psychology of the speaker. The embellishment is made because simple one-one correlations between sensory input and behavioural output would not require the postulation of mental states as intervening variables.

3. Nothing, therefore this neither.

4. To borrow an example from Elizabeth Anscombe [31].

5. This condition of hypotheticality is perhaps not clearly expressed. It is meant to rule out of the class of theoretical entities things like pistons which qualify by the first condition.

   It might be objected that it is wrong to characterise hypotheticality as a condition of theoreticality - there seems to be a sense in which electrons are still theoretical entities even now that their existence is more or less accepted fact. In reply, I would say that I do not mind if "theoretical" is sometimes used in this way, although I would insist that there is also a use in which hypotheticality is a condition of theoreticality. The desire to call electrons theoretical has a number of sources: the fact that fairly recently they were hypothetical entities (and therefore theoretical in my sense) is one; another is the conflation of the theoretical/non-theoretical distinction with the non-observable/observable distinction, together with the (now dubious) assumption that electrons are non-observable; a third is the fact that electrons are mentioned in theories, although this is of questionable significance since the word "electron" only appears in discursive appendices to mathematically-
expressed theories and in explanations-for-the-layperson. Whatever the merits of the usage, the fact remains that Rorty is committed to the hypotheticality of mental states, and so I shall have refuted his position if I can show conclusively that mental states are not hypothetical.

6. I cannot offer a strict definition of what I mean by "primary function" and "primary and pervasive function". There are, however, a number of ideas which I am trying to capture in the claim that the primary function of mental language is in causal explanation: (i) most uses of mental ascriptions serve the function of causal explanation, (ii) the appropriate way to justify most ascriptions of mental states is in terms of their success in accounting causally for behaviour, and (iii) the main reason for regarding mental states as existing in the world is the success of reference to them in causal explanation. I am particularly concerned with questioning (iii) because I believe that the existence of mental states is guaranteed by the non-causal functions of mental language. If this is so, then Rorty is wrong in representing the issue of whether mental states exist as largely a scientific matter. I turn to this in Chapter 6.

7. Sometimes my B-ing is explained in terms of mental states whose object is my A-ing, when B-ing is an accidental result of A-ing. But here the explanation works via my A-ing, with an added hypothesis to the effect that in these circumstances A-ing involves B-ing.

8. In some cases, e.g. in the activity of dance, the object may be a physical movement pure and simple. On my view, this would be treated, I suppose, as the limiting case of an action. What I am concerned to deny is that in every case the objects of my intentions are physical movements.

9. We sometimes speak of movements causing the actions which they constitute, especially in cases of inadvertence. E.g., "In twisting round suddenly he dislodged and broke the vase" might also be expressed as "His twisting round suddenly caused him to dislodge and break the vase". Notice that this kind of locution is out of place when the action inserted is the intended action (e.g. "His twisting round suddenly caused him to pirouette"). In both these cases, furthermore, it is an action which is explained ultimately, not a movement, as the view under discussion assumes, and they are therefore not counter-examples to my present thesis.

10. The position I recommend here draws heavily upon the work
of Donald Davidson [33].

11. I use "desire" here as a general term for all sorts of volitional attitudes, such as those captured by "wish", "hope for", "long for". The attribution of desire, in a more normal sense, may not be appropriate in all cases, but this fact should not be taken as inconsistent with my present claim.

12. Again I ride roughshod over the subtlety of ordinary language by using "belief" and its cognates as general terms to cover all sorts of doxastic attitudes such as those captured by "know", "fear that", "hope that", "calculate that".

13. This is not the place to develop a moral and ethical theory, but lest the preceding remarks be misunderstood I should supply, by way of a few riders, a sketch of the one I hold. I do not think the indeterminacy of rationality infests moral judgment all the way through. Specific questions about specific actions are most often settled, in determinate fashion, by finding out if some course of action is sufficiently supported by some set of moral reasons. (There are questions, but also widespread agreement, about what counts as sufficient and what as a moral reason.) The indeterminacy creeps in only in the area of the evaluation of general moral principles and of the justificatory relations between these. Sometimes an indeterminacy at this level will show through in some specific moral dispute, such as the recent one about abortion.
Chapter 5: Hypotheticality and the Reporting Role of Mental Language

(a) Introduction

Having begun my argument against the claim that mental states are theoretical in the sense expressed by Thesis II, I turn to the issue of whether they are such in the sense expressed by Thesis III. In arguing that they are not I first consider a poor reason for believing my conclusion - the Myth of the Given - against which Rorty's view can sometimes be seen as directed, and successfully so. My own reasons for believing that Thesis III is false stem from the criterial view of mental language suggested by Wittgenstein, and it would perhaps not be relevant to mention them in full here. What I shall attempt to do, then, is to show that Rorty's attempt to justify Thesis III does not work, in so far as his view cannot account for the phenomena of mental discourse, most notably the incorrigibility of certain mental reports.

It is a consequence of Rorty's version of eliminative materialism that as far as we know mental states may not exist. This is what I mean by the claim that mental states are hypothetical (this notion thus having nothing to do with the alleged dispositional nature of many mental states). I am
treating the claimed hypotheticality of mental states as one aspect of their claimed theoreticality because I believe this approach offers the deepest insights into the motivation to deny the existence of mental states which is embodied in Rorty's view.

It will be remembered from Chapter 2, section (b), that on the contingent identity theory the hypotheticality of mental states is not a feature of their theoreticality. This view, in its more sophisticated versions, has a rather special account of theoreticality, at least for mental states, according to which it is a result of the meaning of mental terms that they refer to whatever mediates perception and behaviour (in certain describable ways). Hence on the assumption that something has this causal property, the existence of the mental is guaranteed in advance, as it were, of the scientific discoveries which lock it into our view of the world. This view of theoreticality is not, however, shared by Rorty, for on his view there are competing theories for the explanation of behaviour, mind-theory being merely one of these, and a poor one at that. In what follows I shall use "theoretical" in the sense adhered to by Rorty, in which theoreticality implies hypotheticality.¹

It seems to me that mental states are not hypothetical, that we already know, so to speak, further than their possible
non-existence. This is the thesis of the present chapter.
Now there is both a right way and a wrong way of arguing to this conclusion. What I propose to do is to display the wrong way, to show how Rorty's view is a justified reaction to this aberration which yet is founded on the same fundamental mistake, and then to pursue the right way in the light of the preceding diagnosis.
(b) Phenomenism and the Myth of the Given

For want of a better name I shall refer to the wrong way as "phenomenism". The view has its crudest expression in the intuition that our knowledge of the existence of sensations (and other mental states) is grounded in their phenomenal immediacy - they as it were hit you in the face. Those who have this intuition might argue against materialism by pinching themselves (and possibly others too) and appealing to the evidence of introspection. However, this version of phenomenism has little to recommend it. Severe headaches do not have more existence than mild ones, nor is one more sure of their existence. To put the objection less facetiously, the issue is not whether something can be known to be going on when I pinch myself, but whether that something is correctly described as pain.

A more principled version of phenomenism is expressed in terms of the alleged incorrigibility of first-person, present-tense sensation-ascriptions. Provided I speak the language correctly, I simply can't be wrong about whether (for instance) I am in pain, nor therefore about whether pains exist - such, at any rate, is the traditional view. I have already (in Chapter 3) given my reasons for questioning this
version of the incorrigibility-thesis. However, even if ultimately the thesis is correct, it should be clear that in this context it is question-begging to insist upon an incorrigibility whose existence is denied by Rorty. Moreover, the same objection can be made to the sophisticated as to the crude version of phenomenism: the issue is not whether I have knowledge, but whether whatever knowledge I have is correctly expressed by the self-ascription of sensations such as pain.\textsuperscript{3}

Phenomenism is a version of the Myth of the Given. As such it provides much fuel for the attempts of post-Sellarsian philosophers to attack this Myth, and thus, I now hope to show, is responsible for the plausibility of some version of eliminative materialism, which to many appears to be the only alternative.

It is perhaps impossible to state the Myth of the Given without appeal to metaphor. It is something like the belief that there is a set of things whose existence is obvious and undeniable, which would have to be admitted into every ontological accounting of the world. What this comes to is perhaps best understood by examining the arguments in favour of views which can be described as versions of the Myth. Perhaps Strawson's [36] neo-Kantian arguments for the basic role of material-object-concepts in our conceptual scheme can be construed as allotting to material objects the privileged
office of the Given. 4

Most versions of the Myth, however, place in the role of the Given sensations and other phenomenal occurrences. The latter are ideal candidates, especially if one grants some Cartesian beliefs about the mind and about knowledge, and the phenomenism which these involve. For, it is held, my knowledge of my own sensations is incorrigible and it is out of this that I construct (or posit) my beliefs about everything else, so that if knowledge of anything is possible, knowledge of sensations is also - it is a condition of all other knowledge. It is characteristic of the Cartesian view that no distinction is made between sensations and other mental states, the latter are treated on the model of the former, and hence the view has arisen that what is Given, to a person at a time, is the knowledge of his or her mental states at that time.

This last statement is the epistemological formulation of one version of the Myth; but also involved in the Myth is the belief that this statement has an ontological consequence, namely, that mental states must exist. I am not sure whether to take the Myth as implying the necessary existence of the Given, in which case it clearly involves a modal fallacy, or as implying the existence of the Given relative to what we know. The first alternative is false, the second trivial, or
so it seems. The only way I can think of to invest the Myth with content, while preserving its *prima facie* plausibility in the face of this dilemma, is to take it as involving the supposition that there are certain facts which we **cannot help but know**, the constituents of which therefore we **cannot but admit exist**.

The above formulation of the Myth of the Given gets us a step farther, but leaves many questions open (Who are *we*? Under what **conditions** can't we avoid knowledge of certain facts?) - so many questions, indeed, that I doubt whether a coherent notion of the Given is even available to us. Here I agree with many eliminative materialists and other post-Sellarsians that the notion of the Given is a suitable object for attack. Where I strongly disagree is in the matter of whether the death of the Given involves the abolition of mentality. I claim not.

I am inclined to suppose that Rorty, in his eliminative papers, misconstrued the present problematic by assuming that the only alternative to materialism is phenomenism, and hence that the only reason for believing in the existence of the mental is a thoroughly bad one. Roughly speaking, he accepts the identification, implied by phenomenism, of the Given with the mental, taking this to be part and parcel of the ordinary concepts of the mind and of mental states.
The justification of the attribution to Rorty of the above confusion is by no means easy, since I would not expect him to avow the position openly, even if I am right in thinking that it underlies his general view. We can note in support of this motivational hypothesis the following points:

(i) Rorty's main dissatisfaction with mental states is their alleged irredeemably Cartesian character, the alleged incorrigibility of first-person, present-tense ascriptions; he argues, furthermore, that this incorrigibility extends, though in a derivative sense, not only over the familiar "occurrent" states, but also over those commonly thought of as "dispositional". My treatment of his remarks about the Cartesian nature of the mind has so far been concerned only with the manner in which the supposition that mental states are incorrigible has provided grounds for denying in advance the plausibility of any psycho-physical identity theory. Now, however, more must be said, for those features of the mental which Rorty marks out as the distinctive ones, and therefore the ones responsible, as it were, for the non-existence of the mental in the face of the theoretical superiority of neurophysiology, are precisely those features which lend plausibility to phenomenism, the identification of the mental with the Given. Phenomenism involves the view that mental states are states of which one has incorrigible knowledge;
even the tendency of phenomenism to treat all mental states as akin to sensations has its parallel in the claim that incorrigibility extends to dispositional states.

(ii) In his later work Rorty sometimes writes as if the main object of attack were the Myth of the Given, as a general epistemological/ontological thesis, the philosophy of mind providing only a context for this general debate. The supposition that he sees a clear distinction between the issue of the Given and that of mentality leaves it as at most fortuitous that the philosophy of mind provide the relevant context, for of course if the Myth is false, then at least mental states are no worse off than anything else, for no entity is such that we cannot but admit its existence. (Every entity can be finessed.) The fact that he thinks there is something especially wrong with mental states is then evidence for my motivational hypothesis.

It cannot be sufficiently emphasised that this motivational hypothesis is no more than just that, and that its usefulness is thus limited. I wish to stress that I am not asserting that Rorty's version of eliminative materialism, as a thesis, implies the identification of the Given with the mental, but a view can be based upon a mistake without implying that mistake, and it may still be of interest to expose that mistake.
(iii) Sellars [17], from whom Rorty derives the theoreticalist aspect of his view, explicitly distinguishes the Myth of the Given from the view that there are no mental states (or, in the context of his debate, sense-data), but makes a different error to emerge with the same false dilemma between phenomenism and theoreticalism. The error is to lean too heavily on the distinction between "theoretical" and "observable". Sellars appears to equate the thesis that mental states are observable with phenomenism, with the consequence, typically Cartesian, that mental states are really observable only by their possessors. Sellars is correct to object that this view leaves us unable to account for the intersubjectivity of mental language, the fact that to all appearances we experience no difficulty, in ordinary life, in ascribing mental states to others and in testing, confirming and disconfirming such ascriptions. His conclusion is that mental states are not observable and therefore are theoretical, the Myth of Jones providing a plausible scenario for their introduction into discourse. While the initial tenor of his remarks might lead one to suspect that he is only providing a way to see that there is an alternative to phenomenism, his final comments indicate that he sees theoreticalism as the only true alternative and indeed as the only true account of the mental.
I say that the reliance on the theoretical/observable distinction is an error because the sense of "observable" to which Sellars appeals is the empiricist, positivistic one in which only present mental states can be observed, and these only by their possessors. In this sense tables are unobservables. It does not follow from this, however, that tables are theoretical entities, posited for the causal explanation of some as yet unspecified tabular effects. Nor does the corresponding conclusion follow in the case of mental states. (My own position is, moreover, that the mental states of others are observable (in a very good sense, though not the one above). It remains to be seen whether this form of observability has any consequences vis-a-vis their theoreticality.)

To return to my motivational hypothesis about Rorty, we have seen that the published views of both Sellars and Rorty involve the supposition that the only real alternative to their theoreticalism (of which, admittedly, they hold slightly different brands\(^1\)) is phenomenism. This is the epistemological parallel of the ontological thesis, also held by both (though not central to the topic of this essay), that the only alternative to materialism is Cartesianism. My position is that while materialism/theoreticalism has derived much of its support from its success as an alternative to
Cartesianism/phenomenism, and has provided strong arguments against the latter, there is a tertia via. This can be derived, I argue in the Appendix to this essay, from the criterial view of mental language.

I claimed earlier that both phenomenism and Rorty's eliminative materialism rested on the same fundamental mistake. What I meant by this is that they both accept a certain conception of mentality, according to which anything mental has a certain feature, namely the office of being the subject of an incorrigible report. Phenomenism says there are things, namely mental states, which have this feature. Rorty says that it is possible that there are no things which have that feature, but that the concepts of the mind and of mental states imply that mental states have it, for which reason it is possible that there are no mental states. My own view, part of which I have outlined in Chapter 3, is that there are things, namely mental states, which have been thought to have that feature, but which in fact don't. (Actually some are the proper subjects of reports which, though not incorrigible in the traditional, logical sense, are so in a very different (criterial) sense, to be discussed in section (e) below.)

It is by a return to the notions of report and of incorrigibility that I begin my attack on the hypotheticality
aspect of theoreticalism. It will be remembered that Rorty
[10] supposes that the replacement of mental talk involves the
taking over both of its explanatory and of its reporting, or
observational, functions. While I agree with him that the
phenomenist account of the latter, according to which certain
mental reports are logically incorrigible, with the attendant
ontological consequences, is wrong, I cannot agree that his
own account of mental reporting, which appears to allow an
easy progression into the key of neurophysiology, is adequate.
(c) Mental Reports

A discussion of the alleged incorrigibility of mental reports should begin with the notion of report. What I shall mean by "report" is any utterance which is used to assert (or to attempt to assert) the existence of some particular state-of-affairs, or alleged state-of-affairs. The class of reports is thus a proper subset of the class of constative utterances. What I shall mean by "mental report" is a report of a mental state (or event etc.), where this means no more than a report ascribing a mental state to a person, or other being, and expressed in mental language. (I do not say "report on a mental state", for as this phrase is commonly used there would be no such reports if there were no mental states. I wish to consider the analysis of a certain class of utterances, without prejudice to the question of whether there are mental states.)

The above definition of "report" allows for the existence of theoretical as well as non-theoretical reports. I do not think it worth attempting a rigorous definition of the contrast I have in mind here, for such a definition would contain elements which were heavily relative to the context of discourse, but a pair of examples may suffice. Reports of
quarks made by physicists on the evidence of observations of cloud-chambers are theoretical; reports of baseball scores are non-theoretical. It seems to me that over time theoretical reports come to be non-theoretical reports, as the entities whose existence is thereby reported acquire an accepted ontological status, but these are matters which are best dealt with elsewhere. For the moment I wish simply that my terminology be neutral on the question of whether mental entities are or are not theoretical.

It appears to be a consequence of Rorty's view that it is possible that all mental reports are false, since it is held to be possible that no mental states exist. Strangely enough, this consequence is resisted by Rorty and other eliminative materialists on the grounds that it is too counter-intuitive. It is indeed counter-intuitive: consider two men, David and Michael, the former of whom has, and the latter of whom has not, as we should ordinarily say, a headache. Now even if the ascription of headaches has horrendous Cartesian implications which we should wish to deny, there seems to be some sense in which we are right when we say "David has a headache" and wrong when we say "Michael has a headache". According to Rorty the reason for this is that David is in some neural state in which Michael is not, and that we misdescribe or misidentify his being in this state.
as or with his having a headache, with all its mentalistic overtones. There is a sense in which mental reports are reports about neural states: this allows Rorty to admit that there can be true sensation-reports even though there are no sensations (and, more generally, true mental reports even though no mental states). ¹²

However, having saved us from the fate of believing a large number of falsities (ordinary ascriptions of mental states), Rorty runs into further difficulties, for the notion of report which he is presently employing seems to differ rather a lot from any familiar notion. We should normally suppose that the following principle, which I call the reference-principle, is true:

(R): It is a condition of X's being a report about Y that either (i) X contain an expression referring to Y, ¹³ or (ii) the utterance of X imply some further proposition, Z, which contains a term referring to Y. ¹⁴

This condition is not fulfilled in the present case. Mental reports do not contain expressions referring ¹³ to neural states, nor do they imply any propositions about neural states. In what sense, then, are they reports about neural
states?

A possible answer to this question has been offered by Shope [36]. We should note first of all that the above condition is only plausible as a condition of X's being a report of Y. There is also the expression "X is a report on Y" which operates under no such constraint. To employ Shope's examples, a person may give a report of a piece of play-acting which is in fact a report on a murder. Similarly, a report of a break-in may be merely a report on a householder's regaining entry to her house having lost her key. The recourse for Rorty is then to say that reports of mental states are reports on neural states.

Unfortunately this grammatical sanctuary provides shelter only at a further cost. In most ordinary cases where someone makes an (apparently false) report of (a) Y, which turns out to be a (true) report on some W, the following principle, which I call the belief-principle, seems plausible:

(B): If A makes a report (X) of a Y which is also a report on some W, then A believes, of W, that it is a Y, or, in cases of deception, intends his/her hearer to assume that s/he has that belief.

(Note that I do not say that s/he believes that W was a Y, for
s/he may have no de dicto beliefs about W.) Now it seems a bit odd to say of all of us that we believe, for example of the firing of our C-fibres, that it is a pain, even when we report being in pain.

This, however, is an intuition which I do not expect to share with the eliminative materialist; the reader will make up her/his own mind, but I shall pass over this objection in favour of one more telling. The eliminative materialist account gets into trouble on account of what I call the existence-principle:

(E): If X is a report (true or false) on W, then W exists.

The problem is that the analysis of mental reports as reports on neural states applies only to those reports which are true (judged, as it were, from the ordinary mentalistic perspective). If I report that David has a headache, there is some neural state on which I report, according to the present version of the eliminative view, but if I report that Michael has a headache (which is, as we should ordinarily say, false), there is no relevant neural state to be reported on. (The fact that I can make the general claim to be reporting on Michael's neural states is neither here nor there.) Such a
report is then neither a report of a neural state, because it violates the reference-principle, nor a report on a neural state, because it violates the existence-principle. In that case it is in no ordinary sense a report about a neural state.

The present analysis of "mental report" does not allow, then, for the possibility of there being false reports about neural states (other than those expressed in neural language). This is a severe defect in such an analysis - one would not accept, for example, an analysis of "proposition" which had the result that only true propositions were propositions. The only way for Rorty to avoid this conclusion, while preserving the relevance of the "of/on" distinction suggested by Shope, is to deny that he is trying to supply a full-blown analysis of mental reports, as opposed to a convenient way of dealing with a certain subset of them. We shall see later why this move is unattractive.

I conclude for the moment that the recourse to the notion of "reporting on" is fruitless. I wish now to mention another account of reporting which might be claimed to be relevant here. That is some form of causal account, in which what makes X a report about a certain Y is Y's being involved in a certain way in the causation of the utterance of X. While I do not know of anyone who has explicitly held this view in the present context, I feel it worthwhile showing that
it suffers from the same inadequacy as the appeal above to reporting on.

In this context the view can be articulated as follows: what makes a report a report about a brain-state is the fact that that brain-state causes the making of the report. Again I am not sure if the account is intended as an analysis, or part of the analysis, of the notion of report in terms of necessary and sufficient conditions for something's being a report about something. I shall treat it as if it were, but I think my objections to this account would be worth considering even if it is not to be so rigorously construed.

It is fairly plausible to assume that some form of causal connection between report and reported is a necessary condition of the report's being about the reported. Certainly for X to be a report on Y, as opposed to a random statement which happens to be true on account of something about Y, the utterance of X must be caused, in some way, by Y. The same appears to be true of true reports of Y (true qua reports of Y, not qua reports on some Z); however, the case of false reports of some Y raises problems. In this case there is no Y to cause the report, which is yet a report of Y, and so the causal condition is not necessary.

Assessing the sufficiency of the causal condition requires elaboration of the analysis. Clearly, that Y causes
X is not a sufficient condition for X's being a report on Y. Lung-cancer is not a report on cigarette-smoking; the movement of one billiard-ball not a report of the impinging on it of another. X must also fulfill some linguistic or quasi-linguistic criteria for being a report - it is fairly easy to imagine what these might look like without having to display them in detail here. 15 Allowing this, we also have to be careful about how we identify, out of all the causes and causal conditions relevant to the utterance of a report, the cause which is to count as its object (that which it is about). Let us suppose that in the case of mental reports this can be done by restricting the scope of "the cause" to the neural states of the person reported on (whether by her/himself or by another).

Here again, however, false mental reports present us with a problem. Suppose I lie about whether I have a headache in order deliberately to deceive you into thinking I have one when I don't. I have already made the point, against the necessity of the causal condition, that there is no neural state "corresponding to" 16 the headache which causes me to make the report. There is, however, a neural state which causes me to make the report - part of it "corresponds to" my desire to deceive you - and this appears, on the view under consideration, to fulfill all the criteria for being the
object of the report. Yet it is decidedly not what I am reporting in this case, and so the causal condition is not sufficient.

The next resort for Rorty in his attempt to avoid these objections, which question both the necessity and the sufficiency of the existence of a certain kind of causal connection in each particular case of a mental report, is to insist that the causal theory of reporting demands only a general causal connection between report and reported. The suggestion is that what makes a report a report about a neural state (when it is not superficially so) is that it is a report of a certain kind (mental reports, identified linguistically) and that reports of this kind are generally or normally caused by neural states of the same kind. Since the normal reports are here to be identified with those which we should normally accept as true, this supposition relies on the assumption that most mental reports are true, in the derivative sense here relevant. This assumption may or may not be true - I do not think it can be established by transcendental argument, as some have thought, but that is another matter.

My objection to this form of the causal theory requires raising stones hitherto left unturned. Since the theory makes no distinction between true and false reports, as regards what being a report about something amounts to, it seems sufficient
to note, against the necessity of the above weakened condition, that there are reports about unicorns though no unicorns to cause them. Against its sufficiency I shall note that there are many other general factors in the causation of mental reports, such as the presence of oxygen at the surface of the planet, the genetic endowment of reporters and others of their neural states, which are not part of what is reported by any mental report.

It seems clear that these objections apply also to the particularised causal theory, taken as a theory of reporting tout court. What emerges from these considerations, then, is that while the causal account may be plausible as an analysis of "report on", it cannot work for "report of". On account of the existence-principle therefore it applies only to (ordinarily) true mental reports and cannot be regarded as a comprehensive account of reporting.

What this means is that I must consider the possibility that the accounts criticised above are not intended as comprehensive analyses of the truth-conditions of mental reports, but merely as descriptions of ways in which we can sometimes admit that one who makes a mental report says something true even though his/her utterance is strictly false. What I am supposing is that the latter alternative does not commit one to any account of false mental reports,
which, if it were the case, would disarm my objections both to the causal account in general and to the relevance of the "report of/report on" distinction. This alternative would stress the analogies between the case of mental reports and those, to which Donnellan [40] has drawn our attention, which provided the original motivation for causal theories of reference and reporting. If, for example, I say that the man with the martini is the Lord Mayor, I may succeed in making a true report about him even if there is only water in the martini-glass.

The first problem with this strategy is that it is of course plausible only for those mental reports which contain (what at least purport to be) definite descriptions of mental states. One cannot salvage any vestige of truth for "David has a headache" by these means, any more than one can for "John kicked a man with a martini", where John did kick the Lord Mayor. So if the eliminative materialist adopts this strategy instead of the one giving a full-blown analysis of mental reporting, the scope of his/her account is considerably smaller than intuition demands.

It is by no means clear that this alternative strategy would be attractive for Rorty even for those reports which are most closely analogous to the Donnellan examples. Consider the following case: I observe Patrick and Tony each in an
agitated condition, the former more so than the latter. I explain this situation by reporting that Patrick's indignation (about some unknown state-of-affairs) is more intense than Tony's. Let us suppose, however, that the cause of their agitation is not indignation (nor any neural state apt to be misidentified as indignation), but indigestion. Now there is some sense in which my report about relative intensities is true of their states of indigestion, or at least there is some plausibility for this claim. The point I wish to make is that on the present eliminative materialist strategy there is no significant difference between the above case and one in which the true cause is supposed to be a neural state.

While I have nothing to say directly against this conclusion, it seems to me one that would be unattractive to Rorty. Although, to my knowledge, this issue has not been confronted by him directly, I am inclined to suppose that in his view the truth of certain mental reports is to be allowed as a result of the fact that they are reports in the language of a theory which is to be replaced and that they can be reinterpreted as reports about the entities referred to in the replacing theory. The scenario for theoretical change does not apply to the indigestion-example. For this reason I suggest that the required account of mental reports is something more than a collection of Donnellanesque
observations, that it is indeed intended as a comprehensive analysis of mental reporting, and that therefore my original objections are still relevant.

I am afraid I can think of no way of amending or complementing the causal analysis of reporting which would evade the above objections, short of relying on some form of the reference-principle. But to do this would disallow neural states from being the objects of mental reports and so undercut Rorty's account of these reports. It is perhaps significant that the difficult cases for this account are those involving false reports; one can only assume that these were overlooked because the main topic of enquiry in this field has always been the so-called incorrigible reports, and it has been assumed that these are always true. But an analysis which only works for sincere first-person present-tense mental reports is not an analysis of "mental report".

It might be argued that my wide definition of "report", which allows for "theoretical", or hypothetical, reports, is the real culprit here, for the causal analysis was developed only as an account of so-called "observational" reports. This counter-objection, however, does not show enough. Even if it were correct that second- and third-person mental ascriptions are not strictly reports, because they consist in the making of (causal) hypotheses, there remain the difficult cases of
(deliberately) false first-person reports which are by themselves capable of defeating the causal analysis. Moreover, to be consistent with Rorty's theoreticalism, the causal analysis must allow for such hypothetical reports.

It does not seem plausible, then, to regard the causal account as helpful to Rorty. It is possible, however, that an account of reporting can be constructed from his observations about reference [39]. Here Rorty distinguishes three senses of "refer", and while I do not believe that this approach is ultimately successful, I think it is important to consider it as a possibility for the expression of which Rorty has given us the basic conceptual tools.

The tripartite distinction is roughly as follows: referring₁ is the "commonsense" notion in which the referrent of an expression is determined by the speaker's intentions; referring₃ is a "philosophical" notion to which the existence-principle but not the reference-principle applies—remember that the sense of "refer" in the reference-principle is "refer₁". Again roughly, we can say that referring₁ is involved in reporting of, and referring₃ in reporting on. The key notion for Rorty is referring₂. This is "an intermediate notion" in which "one can only talk about what exists, but in which the truth of one's remarks is not determined by the discovery of what one is talking about." When we say (truly)
that David has a headache, we are referring₂ to the neural state "corresponding to" his pain, but to ascertain the truth of what we say, we do not need to track down the relevant referent by causal or other means (this distinguishes referring₂ from referring₃).

While the notion of referring₂ allows Rorty to sidestep the debates about whether reference is causally or otherwise determined, and about whether determining truth requires referential access, it is inadequate as a recourse in the analysis of mental reporting. If mental reports involve referring₂, then by Rorty's definition of this notion they are subject to the existence-principle. The analysis is then open to the same counter-examples (ordinarily false mental reports) as were responsible for the demise of the "reporting on" and causal accounts. The difference between referring₂ and referring₃ is not the difference relevant to these counter-examples.

The results of this section can be summarised using the tripartite distinction of referring. Shope's suggestion, and the causal analysis discussed after it, are attempts to analyse reference₁ to mental states in terms of reference₃ to neural states. These founder on the existence of (ordinarily) false mental reports. Reference₂ can be brought in to avoid certain objections to the programme of accounting for
reference₁ in terms of reference₃ - a task which Rorty [39] does not think is possible. However, reference₂ fares no better: the same counter-examples apply because the same constraint upon the determination of reference, namely the existence-principle, applies.

In the next section I move to some further arguments against the general theoreticalist analysis of mental reporting which Rorty inherits from Sellars. Here I look not at false mental reports but at true, because incorrigible, ones.
(d) Incorrigibility

In this context, the theoreticalist analysis of reporting is part of an attempt to overcome what appears to be a major difficulty for any theoreticalist view of the role of mental language. For if the primary characterisation of that role is in terms of the making of causal hypotheses to explain overt behaviour there is a problem about how first-person present-tense mental reports are even possible, given that they are not in any way based upon the observation of the subject's behaviour. This feature of first-person present-tense mental reports is usually called their non-inferential nature. It is to be distinguished from their alleged incorrigible character, which brings problems of its own, to be discussed later.

The Sellarsian account [17], adopted with modifications by Rorty ([11] p.411ff.), of the possibility of non-inferential reports consists in a story about how we are trained (or could be trained?) to make them. Suppose Jones the caveman observes Smith exhibiting the kind of behaviour which, on Jones' mind-theory, leads to the hypothesis that Smith is in pain. Once the theory has been publicised and generally accepted, everybody else knows that it is appropriate to say here "Smith is in pain". In this situation, Smith's being in
pain provides us with an ideal opportunity to begin training him to utter the words "I am in pain" when he is; this training will of course take place over a span of time, involving a number of occasions when Smith is and is not in pain, with suitable rewards and punishments for uttering the self-ascription. Eventually, and this is what constitutes successful training, we will have established a direct causal connection between the hypothesised cause of his prior grimacing (the pain) and his mental report. The following diagrams chart the causal situations before and after the completion of training, where "p" = "Smith is in pain":

**Before**

1. Smith's pain
2. Smith's grimace
3. Others say "p" (inferential)
4. Smith says "p" (inferential)

**After**

1. Smith's pain
2. Smith's grimace
3. +Effect of Training
4. Smith says "p" (non-inferential)

---

**Figure 1**
In the eliminative materialist modification of this picture, the initial causal hypothesis is incorrect, and "Smith's pain" must be replaced by the actual cause of his behaviour, which is some neural state N. The criterion of success for the training is now that this brings it about that N causes Smith to say "I am in pain" without the intervention of other people's reports. (It is not of course a feature of either of these pictures that the initial cause always results in a report, even given correct training - other factors can always block the disposition to utter a self-ascription.) On any particular occasion after the completion of training, Smith's report will be non-inferential in so far as it is not caused by any process of inference. Smith himself need not even know the theory which at first led others to train him to make such reports. Nor is what he says the effect of any immediate process of inference by other people, as it would be in the early stages of training, when they reward him for imitating their reports, which are inferential.

This story, in both its mentalist and materialist forms, does allow for the possibility of non-inferential mental reports, in the sense mentioned above. It is even a plausible account of how we come to make mental self-ascriptions. I believe there are difficulties for the story, however, when the notion of incorrigibility is introduced.
It is obvious that the theoreticalist account cannot support any form of logical incorrigibility. That Smith utters the sentence "I am in pain" (sincerely and in a state of linguistic competence, so to speak, does not entail that he is in pain. It is easier to show this on Rorty's version, when of course the truth is not that Smith is in pain, but that he is in neural state N. Even on the more general version, however, the conclusion is the same. The fact that I have been trained to make a certain response in certain kinds of situation, and to inhibit that response when the relevant kind of situation is lacking, does not entail that whenever I make that response the situation must hold. One could try to make this denied entailment a necessary condition of linguistic competence, but that would make the thesis that sincere linguistically-competent mental self-ascriptions are incorrigible into a tautology, since never making a mistake would be written into the definition of competence. While the belief in logical incorrigibility might be the result of being overly impressed by some tautology of this form, those who hold the belief usually insist on a certain amount of distance between linguistic competence and the impossibility of error.

In any case, since this essay is part of an argument against Rorty's version of eliminative materialism, I can afford to concentrate on the eliminative version, with a nod
of acknowledgement towards the more general (or mentalist) version. It is not, in my opinion, a fault in Rorty's view that it does not support logical incorrigibility, because I am not convinced that there is such a feature of mental, or even of phenomenal, states. What I am convinced of is that there is a good reason for insisting on a certain kind of epistemic priority for, in this context, non-inferential reports over inferential ones. However, this form of epistemic priority will turn out to be much less significant than has usually been thought, and also to be not that in which consists whatever form of incorrigibility is attributable to non-inferential reports of mental states.

These somewhat cryptic remarks will become clearer through a consideration of Rorty's account of incorrigibility. According to this, incorrigibility consists in the fact that, for a certain reason, non-inferential reports are not challenged by other people. This "empirical" form of incorrigibility would easily evaporate if the relevant challenges were to take place and were to lead to the overthrowing of non-inferential reports. Its existence is a fact of sociology, not of metaphysics or epistemology. As Rorty defines it:

\[ S \text{ believes incorrigibly that } p \text{ at } t \text{ if and only if:} \]

(i) \( S \) believes that \( p \text{ at } t \),
(ii) There are no accepted procedures by applying which it would be rational to come to believe that $\neg p$, given S's belief that $p$ at $t$. ([11] p.417)

According to this view, then, incorrigibility consists in a form of epistemic priority allotted to non-inferential reports over inferential ones, although the kind of priority is considerably watered-down compared to the kind associated with logical incorrigibility. For Rorty it has the advantage of involving no ontological commitment to mental states.

The bare sociological fact that certain sorts of reports are not challenged and not overthrown does not seem sufficient as an account of incorrigibility. If reports by red-headed people suddenly became sacrosanct within some society, this would be of little philosophical interest, though perhaps it might have great sociological and psychological implications. What turns the social fact into a philosophical point is the existence of a good reason why such reports are not to be challenged. In the mental case, what we are looking for is a good reason for allotting a certain kind of epistemic priority to non-inferential reports over inferential ones, and Rorty attempts to provide one.

Two non-starters should be mentioned first. We cannot say that the epistemic priority of non-inferential reports resides in the possession of better epistemic credentials.
This is a view associated with the Myth of the Given, that non-inferential reporting is done on the evidence of the phenomenon itself, whereas inferential reports employ behavioural, and therefore mediate, evidence. On Rorty's eliminative materialist account, non-inferential reports are made on no evidence, and their justification is only as good as the justification of the inferential reports which are involved in the conditioning process which implants the ability to make them. On this view the epistemic priority allotted to non-inferential reports is a function, not of their having sounder epistemological foundations, but of their being better indicators of the subject's neural states than the other behavioural indicators which could be used to support inferential reports about those states (dressed up as reports of mental states, of course).

The second non-starter arises after this way of viewing reports has been granted: it is the possibility that what "being a better indicator" amounts to is a purely statistical matter, that the reports issue from the neural state more often than does the non-verbal behaviour. What we have to note is that while this may be true, it is not a truth which is available to us at the present stage of the development of neurophysiology and therefore cannot be part of the underpinning of our present conceptions about incorrigibility.
Having disposed of the other possibilities, I turn now to what Rorty actually says. This is that:

... when the behavioral evidence for what Smith was thinking about conflicted with Smith's own report of what he was thinking about, a more adequate account of the sum of Smith's behavior could be obtained by relying on Smith's report than by relying on the behavioral evidence. ([11] p.416)

I wish to argue that this explanation is either irrelevant or incoherent.

It is irrelevant if the range of behavioural evidence is construed too narrowly. It seems that what Rorty has in mind is the behaviour carried out by Smith at the time of his report, or close thereto:

Thus, for example, if Smith's cave re-entering and ax-grasping behavior seemed to point to his just having had the thought that he had left his ax in the cave, his subsequent use of the ax nevertheless confirmed the truth of his report that what he had actually thought at the moment in question was that he might have broken the ax-handle yesterday. ([11] p.416)

In terms of our pain example, the relevant contrast might be between Smith's reporting his being in pain and his failing to wince or grimace.

What is strange about these examples is that they do
not seem to be cases of conflict between Smith's report and the behavioural evidence. In the first case, Smith's "cave re-entering and ax-grasping behavior" is in no way inconsistent with his having had the thought which he in fact reports. What is perhaps inconsistent with the latter state-of-affairs is that whose existence is affirmed by the inferential report which identifies his thought (at that time) incorrectly (as the thought that he had left his axe in the cave), but this shows that the inferential judgment was hasty, based on too narrow a sampling of Smith's behaviour.

Similarly, in the pain case, Smith's failure to wince is consistent with the truth of his report that he is in pain and by itself no more indicates to us that he is not in pain than his failure to signal the fact in semaphore. It seems to me to be a general point about mental states (or neural states, as we must say in this context) that there are vast numbers of manifestations of any particular state, and a vast number of states which can play a role in the explanation of any particular manifestation (sequence of behaviour), so that the above-mentioned failure of evidential conflict will be quite widespread.

A conflict between non-inferential report and other behavioural evidence would occur if Smith exhibited little or no behaviour of the type normally explained by the mental
ascription made in his report. If he neither grimaced, nor
winced, nor moaned, nor held any part of his body, nor called
for an ambulance, nor stole morphine, nor . . . , but reported
being in pain, we do appear to have a case of conflict.
However, here there is little or no behaviour to be explained
by the supposition that he is in pain (save his report), and
therefore no sense can be attached to the supposition that
taking his word leads us to a more adequate explanation of the
sum of his behaviour. In this case, the explanation offered
by Rorty of why we regard non-inferential reports as
incorrigible is incoherent.

The eliminative materialist might retreat to a weaker
position, and this may have been the position Rorty was
implicitly arguing for, namely that the privileged status
given to non-inferential reports is a result of the fact that
when all the evidence for a person's mental (or neural) state
is in, it happens more often than not that the report agrees
with the hypothesis others make to account for that evidence.
On the theoreticalist picture, this concordance in the case of
any particular individual's mental self-ascriptions appears to
be a condition of successful training.

(It should be remembered that all this is true, if it is, only of sincere mental self-ascriptions (for all we know
people may lie about their mental states more often than not),
and so the instances which inductively support the supposed correlation between non-inferential report and hypothesis must be observed in a context where there is also behavioural evidence for the sincerity of the report. On the present view, this would be evidence for the subject's being in some neural state the presence of which we unwittingly report when we attribute to him the belief that what he reports is true.)

This weaker position represents the relationship between behavioural evidence and non-inferential report as one of corroboration as opposed to conflict, at least in general. It does allow for the occasional case of conflict, but with a result which is not desirable for its holder. For if agreement with the behavioural evidence is the (only) reason why non-inferential reports are to be accepted, then disagreement with the behavioural evidence, where this arises, would be a reason for rejecting them. The picture associated with this weaker position gives us no reason whatsoever for preferring the non-inferential report in the case of a clash, and every reason for preferring the hypothesis inferred on the basis of the other behavioural evidence.23

The picture above does give an explanation of why non-inferential reports are accepted without question in a certain sort of situation, namely when very little other evidence is available, for, in view of the general correlation, we are
more likely to be right than wrong in going along with the report. However, if this is what incorrigibility amounts to, it is of little significance to pontificate about the incorrigibility of mental self-ascriptions. For as a matter of fact winces may be even better correlated with other pain-evidence than non-inferential reports, and therefore should be considered incorrigible. If the fact that winces are non-linguistic is an insuperable bar to this, it is not one to the claim that winces are better evidence for pain (or whatever neural state we unwittingly call "pain") and therefore should be relied on in cases of conflict. Furthermore, most ordinary reports about the world are incorrigible in this rather expansive sense, for if no other evidence is in, we do first off accept the testimony of other people, especially if it is attended with the marks of sincerity.

I conclude that Rorty's account of incorrigibility is seriously inadequate. In the main it suffers from internal deficiencies rather than a failure to measure up to the alleged external phenomenon of incorrigibility, but the latter is appealed to in my final rejection of the weaker position above. I shall consider an alternative account of incorrigibility in the next section. Now I must assess the significance of my findings with regard to the present version.
The account of non-inferential reports is not an essential part of the account of mental reports, as this is not an essential part of Rorty's eliminative materialism. It remains open to him to come up with accounts which avoid my objections. I have shown simply that those most commonly associated with the view will not do the job, but this at present is sufficient.

It is worth reminding ourselves of under what constraints any alternative account must operate.

The purpose of the account of non-inferential reports was to give a plausible explanation of their apparent incorrigibility (to the extent that this exists); the biggest constraint here is that both the explanation and interpretation of incorrigibility must be consistent with the non-existence of mental states, the apparent objects of the reports, if the account is to figure as part of Rorty's version of eliminative materialism. (The same stringent requirement also operates as a constraint upon the general account of mental reports.) The above constraint engenders a weak notion of incorrigibility, whereby reports can be incorrigible while lacking any ontological commitment to the entities about which they are, apparently, reports. The account provided (and arguably any account which operates under the non-existence constraint) has the disadvantage that,
when followed through, it allots a privileged evidential position not to non-inferential reports but to the general behavioural evidence available to other people, as I argued earlier. As a result of this the account cannot explain why non-inferential reports cannot be overruled by purely behavioural evidence, why this task must be assigned to the use of cerebroscopes or encephalographs. It seems to me that no account which is based on the assumption of the primacy of non-reportative behavioural evidence can generate a good reason for rejecting the latter on particular occasions in favour of some other form of evidence.
(e) An Alternative Account?

In the argument above I take it as given that the relevant non-inferential reports cannot be overruled by behavioural evidence. This remark is subject to two qualifications: first of all, it is clear that mental self-ascriptions are often overruled on the basis of behavioural evidence when the mental state ascribed is of the "dispositional" kind, what Rorty calls a "mental feature"; this is admitted on all sides. Secondly, and more pertinently, there is the class of ascriptions of mental events (so-called by Rorty) which are overruled on the basis of behavioural evidence. This class is not empty; all that must be noted is that when such an overruling takes place it is followed by one of two judgments, either that the self-ascription was not sincere or that the self-ascriber is, to this degree, linguistically incompetent. In this arguably trivial circumstance lies the heart of incorrigibility. The circumstance is temporary in so far as when we are in a position to acquire it, neurophysiological evidence might allow us to omit either of these judgments.

Now in "Mind-Brain Identity, Privacy and the Categories" Rorty treats incorrigibility in precisely this manner, identifying it with the absence of a distinction between
misjudging and misnaming in the case of first-person present-tense mental reports (the question of sincerity being allowed to rest for the moment). The "Sellarsian" account which up to now I have been criticising is the one drawn from Rorty's other important paper, "Incorrigibility as the Mark of the Mental". It would seem that these accounts require differential treatment, were it not for the fact that Rorty, in the earlier "Categories" paper, treats them as complementary - he moves from what he takes to be the Wittgensteinian account to the theoreticalist one, although the latter is not there developed with as much sophistication as in the later paper. We have already seen that the Wittgensteinian version is represented in the other in the shape of the observation that it is a condition of successful training that non-inferential reports are in accord with the hypotheses generally arrived at by others to explain the behavioural evidence. This means that if we wish to overrule a non-inferential report, we can always attribute it to linguistic incompetence, or so it seems.

On the surface, then, there is a considerable area of agreement between Rorty, myself and Wittgenstein on this matter. We all three oppose the Cartesian account on which incorrigibility is the result of privileged access to a private inner stage, seeing it more as a matter of the
availability and non-availability of certain linguistic moves. We all three agree that neurophysiological evidence could in principle be relevant to the overthrowing of a non-inferential report without it being necessary to charge insincerity or linguistic incompetence. We differ in the explanation of the genesis of incorrigibility, and I have argued that Rorty's Sellarsian account of this is ultimately incoherent.

Wittgenstein's explanation relies on the alleged pragmatic intersubstitutability of expressions such as "I know I am in pain" and "I am in pain" within ordinary discourse ([5] #246). I am not absolutely convinced of the correctness of this move, but will not discuss the matter here. In Chapter 3 I have suggested that a successful account of incorrigibility, which is Wittgensteinian in conception, can be generated from the idea that mental language is governed by criteria.

To return to a consideration of Rorty's view, we should note that there is a discontinuity between the accounts of incorrigibility offered in the two relevant papers, if the former account is as loyal to Wittgenstein as Rorty announces it to be.

For on this account it is a conceptual impossibility that one should be both linguistically competent and sincerely wrong in making an non-inferential report. (It is not a logical impossibility: it is an impossibility which can only
become a possibility as the result of conceptual change.) The notion of linguistic competence employed in the later version of Rorty's view is other than this, and indeed must be so. The later version tells us that we have a good reason, inductively established, for accepting the sincere, non-inferential reports of linguistically competent people. But if the notion of linguistic competence is the same one as in the earlier account, this "advice" is redundant or useless, as there is no way of failing to follow it. This is so because there is no way of identifying linguistic competence in a subject independently of whether we accept all her/his introspective reports. Anything which will lead us to reject her/his sincere report will lead us to reject her/his linguistic competence.

The notion which is required for the statement of the later view is one which makes no appeal to conceptual impossibility. In this sense, the subject can manifest his/her linguistic competence in ascribing mental states to others, or to him/herself on other occasions, in a more or less acceptable way (sc. in a way which accords with the hypotheses generally made by others). In this case, if we wish to overrule a particular sincere mental self-ascription, we do not have available to us the blanket charge of "linguistic incompetence!", since the subject may have
demonstrated his/her competence in countless ways. The main defect in the proposed eliminative materialist account, that despite itself it allots epistemic priority to the general behavioural evidence, is manifested here in the consequence that such an overruling would be possible on the basis of the evidence presently available to us. As a result incorrigibility evaporates.

It may be that Rorty differs from Wittgenstein in not tying the misnaming/misjudging account of incorrigibility to notions of conceptual impossibility. The absence of the distinction between misnaming and misjudging then would not be the absence of a difference between them (in the mental case), but simply that we have no procedures for deciding on which side a false non-inferential report falls, at least until encephalographs are readily available. This removes the discontinuity, but also the possibility of regarding the earlier account as an alternative means of explaining the relevant version of the eliminative materialist position. This leaves us with the later account, against which I have already argued.

When fully developed, the Wittgensteinian account of incorrigibility turns out to be more or less incompatible with Rorty's approach to mental states, which admits the (epistemic) possibility of their non-existence, and which
supposes that mental reports are ultimately reports about brain-states. First of all, it is a consequence of this approach that it is facts about brain-states which make mental reports true or false, and it is difficult to square this with the view that it is conceptually impossible to make a mistake in certain of one's mental reports, namely those which are ultimately reports on one's current brain-states. For the normal criteria for the application of neural-state-ascriptions, where these exist, allow for a clear distinction between misnaming and misjudging, and hence for corrigibility.

A possible recourse in the face of this criticism is to stress the contingency of the relation between mental reports and neural states. We could say that what (it is that) we cannot be mistaken about are not our brain-states (except perhaps in some rather oblique de re way), but something of the form "I am in whatever state has such-and-such properties" (e.g. causal properties). This allows us to state the truth-conditions for mental reports without explicitly referring to neural states, but to admit that the latter may be what make mental reports true or false. The corrigibility of our explicit beliefs about neural states is then no bar to the incorrigibility of our first-person mental reports, since we can know the truth-conditions of the latter to be fulfilled without knowing any truths formulated in the language of
neurophysiology. The only problem with this recourse is that it is not available to an eliminative materialist, since it involves stepping back into the contingent identity theory, as formulated, for example, by Lewis [14]. The recourse is also inconsistent with the claimed hypotheticality of mental states (see Thesis III), since the open-ended nature of the schematic account of the truth-conditions of mental reports brings it about that if anything causes behaviour (in a systematic way) mental states do, and therefore exist.

In general, the attempt by Rorty to preserve our ordinary intuitions about the truth and falsity of mental reports is always in danger of causing his view to collapse into the contingent identity theory, as the claim that neural states make mental reports true and false is always in danger of collapsing into the claim that the truth-conditions for mental reports are to be laid out in terms of neural states. The latter claim sounds like the non-contingent identity theory; the above recourse rescues Rorty from this unpalatable conclusion but carries him only as far as the contingent form of the theory. To get any farther he must rely on the general theory of reports which I have treated earlier on this chapter.

A second incompatibility between Rorty's and Wittgenstein's accounts of incorrigibility surrounds the
notion of epistemic priority. On the latter view, this does not really exist as a consequence of incorrigibility. The fact that we cannot distinguish between misnaming and misjudging in the case of non-inferential reports does not give us a reason for withholding criticism of them. We may be less inclined to charge bad judgment, but to that extent we shall be more inclined to charge linguistic incompetence. (Again, all this may change.) We do allow a certain amount of priority to first-person over third-person reports, but this is mostly because the subject is in a better position, usually, to be able to rule out various possible defeaters of his/her report, such as that s/he is insincere in making it, merely pretending or whatever; and all I mean by "better position" is that we would have to do some work (of observation and possibly manipulative experiment) to find out what s/he already knows. (Of course, sometimes s/he will not know, and sometimes we will already know.)

For reasons such as these, it seems to me that the two accounts of incorrigibility do not complement each other and that we should regard the later, considered version as the one truly representative of Rorty's eliminative materialism. My main objection to this account is that while it is an attempt to justify regarding non-inferential reports as epistemically prior to inferential, third-person reports, the manner in
which this is done leads only to the conclusion that the latter are to be given priority in the case of a conflict. The attempt could perhaps be made in another way, but it seems unlikely to me that it will be successful.

It seems to me, then, that Rorty's view, in so far as it involves a commitment to Thesis III, cannot account for the existence of incorrigibility, as explicated in terms of the criterial view, and that in general his explanation of the reporting role of mental language is inadequate. My next task is to point out that there exist further roles for mental discourse, not covered at all by Rorty's published work, which cause problems for his general view. This project was begun in Chapter 4, where I distinguished reason-explanations from the application of the causal-explanatory role of mental language. It is now time to see if this and other further uses can be accommodated by this version of eliminative materialism.
Footnotes to Chapter 5

1. It might be argued that hypotheticality is not a conceptual condition of theoreticality even on the eliminative view — it's just that mental states are unsuccessful theoretical entities, their lack of success being responsible for their failure to exist. I can allow this. I am concerned with what is definitely a consequence of the view, that as far as we know mental states may not exist, and I do not mind how it is derived.

2. While phenomenism may well be an ingredient in both phenomenalism and phenomenology, there is more to each of the latter than is relevant to the present discussion. The crude version of phenomenism appears to be held by E.S. Shirley [34], who argues, against eliminative materialism, that one cannot eliminate physical suffering merely through linguistic retraining. On this point also see R.K. Shope [35].

3. The argument does not stop here, for it is by no means obvious that one can have knowledge while misdescribing, or that the thesis of incorrigibility allows mistakes of this type. The effect of the objection is to shift emphasis to conceptual matters of this kind (some of which are treated elsewhere in this essay) and away from the relevance of the phenomenal immediacy of sensations etc..

4. There is a matter of interpretation relevant here, for Strawson insists that the basicity of material objects is a feature of our conceptual scheme, not necessarily of all possible schemes. Other philosophers have argued the general claim, and this fact is all I need to create a picture of what it means to appeal to a Given.


7. Rorty [37] p.95, for example.

8. In the words of Sam Coval.

9. Strictly speaking, he is initially concerned only with sense-data, but his later remarks about the philosophy of mind indicate that all mental states receive the same treatment.
10. Sellars himself is not an eliminative materialist, but his theoreticalist views are adopted by Rorty.

11. Admittedly they offer different accounts of what incorrigibility amounts to.

12. Steven Savitt [38] has attacked this question from a different angle. His idea is that a statement such as "Sensations exist" is ambiguous: taken as "internal" to the (theoretical/conceptual) framework of mind-theory this statement is true, in accordance with our ordinary intuitions; taken as "external" to that framework (the distinction derives from Carnap), the statement can be dissented from without damage to our intuitions. (Savitt does not say that on the external rendering the statement is false, since by extrapolation from his application of the distinction to Ptolemaic astronomy the external version is a "recommendation or a practical judgment to use the framework of [mind-theory], . . . and as such is neither true nor false.") This view appears to have the uncomfortable consequence that there are no true or false ontological statements, but that every theoretical existential statement is true (internally, but that is the only way existential statements can be true or false).

13. Rorty [39] has distinguished between referring, referring, and referring. It is the first which is relevant here, for we wish to retain the possibility that there can be (false) reports about something which does not exist.

14. This allows "Quick! . . . outside the bank. . . masked man. . . white horse. . . shot clerk. . . headed west..." to count as the report of a holdup.

15. An interesting dialectic on the plausibility of the causal analysis of reporting can be entered from a consideration of whether or not the turning-red of a piece of litmus-paper is a report on the acidity of the solution in which it is immersed.

16. No theoretical weight is attached to my use of "corresponding to" here - it means something like "apt to be misidentified as".

17. That the causal account is wrong does not, of course, show that the reference-principle is true, but it leaves a large hole in Rorty's account which must be filled if it is to be significant, or important, to use the term "report" in this
philosophical context.


19. If anyone were to insist that Smith's report must in any case be true (on account of its in some sense being a report about his neural state), and therefore is incorrigible in this sense, I would reply that those who argue in this way run a severe risk of having their view collapse into the identity theory, for they have begun to identify the truth-conditions for mental and neural reports.

20. I mean "the sincere first-person mental reports of linguistically-competent people made without evidential reliance on self-observation" over "second- and third-person mental reports" - I shall use "non-inferential" and "inferential" as shorter labels for these types unless confusion would result. It is not primarily non-inferentiality which guarantees epistemic priority. Nor is it obvious to me that reports of the latter class are inferential in any sense other than that explained at the beginning of this section - that of being based evidentially on the subject's behaviour.

21. I ignore the fact that this definition contains primitive mental terms like "belief" and "rational". For our purposes, as in any case Rorty implicitly acknowledges, it would be better to replace "believes" by "sincerely reports". This would avoid admitting the incorrigibility of certain non-mental beliefs (as Rorty shows), avoid any hidden commitment to the existence of beliefs and also preserve the continuity of the present discussion, which is concerned with reports.

22. The notion of consistency being appealed to here is not that of logical consistency, which would trivialise the point being made, but of evidential consistency. P is evidentially consistent with Q iff P is not good evidence for not-Q.

23. Rorty does not consider any case of a non-inferential report being overthrown by behavioural evidence alone. He imagines that this will (can?) only be done using neurophysiological evidence (the results of cerebroscope readings). When he comes to his example of the latter event ([10] p.190), however, he chooses one in which the behavioural evidence goes along with the neural evidence, and both are incompatible with the subject's report. It is hard to know which is here refuting the conjecture.
24. Except in so far as Rorty believes that incorrigibility is the mark of the mental. This aspect of his eliminative materialism is not under scrutiny here.

25. We should have expected this when he (on my elaboration) began to speak of neural states "corresponding to" mental states, for here he was hinting at the existence of certain psychophysical correlations, the assumption of which his view was set up to avoid. See Chapter 2, section (d).

26. My intuitions here can perhaps be supported by reference to another of Rorty's examples: if I wish to interpret the statement "Zeus' thunderbolts cause forest fires" as a true report about lightning-flashes, there seems to be no reason for me to hesitate in identifying Zeus' thunderbolts with lightning-flashes, and therefore no reason for me to insist that Zeus' thunderbolts do not exist.

27. I suppose the very non-inferentiality of most self-reports of occurrent mental states is responsible for a certain degree of epistemic priority, to the extent that, if they are not based upon inference, they cannot be based upon faulty inference.
(a) Non-constative Uses of Mental Language

I begin by looking at some uses of mental language which are not theoretico-causal and asking whether these uses are compatible with the alleged role of mental ascriptions in the explanation of behaviour (sc. physical movements).

By a "constative" use of some sentence I mean a use which states, or which attempts or purports to state, a fact. Such a use need not be a report, in the sense in which an hypothesis, for example, is not a report, for the making of an hypothesis is to count as an attempt to state a fact. I wish to concentrate our attention on those uses of mental language which are not constative in the above sense. I think it is undeniable that there are such uses: what we need are a few examples.

In certain circumstances my utterance of the sentence "I hope he will come" might constitute an attempt\(^1\) to state a fact about myself - for instance if it is part of a reply to my psychiatrist's request to list some of the things I hope for. In other circumstances, however, the utterance may be nothing more than an expression of my hope, if for instance I am alone and anxious about his arrival. Wittgenstein [5]
suggests that we regard such utterances as replacements for
the natural expression of mental states, or perhaps as
developments from these (since the non-linguistic expression
of so specific a hope is somewhat less obviously natural than
the expression of pain through screaming, writhing etc.). He
is perhaps more concerned with dispelling the belief that such
utterances are reports about the events in some private mental
arena, but I think his suggestion has wider application and it
is very possible that this was part of his intention.

To continue with some more examples: the utterance "I
wish you would go away" has the force, in many circumstances,
of an injunction rather than a self-description. "I think
you're lying!" can present a challenge rather than a piece of
information. Certain apparently "mental" constructions can be
used to make tentative judgments about the world: "It looks
(seems) to me as if (that) ...", "I (am inclined to) believe
that ..." are examples of this type of use.

This phenomenon is not confined to, though it is more
commonly found in, first-person utterances. "You expect me to
kneel to you!" is a refusal, not a question. "He deems it
advisable for you to accept" may be the conveying of an
injunction, not an ascription. Impersonal utterances also
have non-constative force, depending on the context and
circumstances of use: "One doesn't put on airs at philosophy
conferences" may be a warning or a threat as opposed to an empirical generalisation.

There are many more examples, but I think we have enough to deal with the philosophical points I wish to make about such utterances. As a preliminary to the discussion, it should be noted that all the sentences I have cited as examples also have constative uses, in different contexts and circumstances. I wish to look first at questions surrounding the relationship between the two types of use, in particular at the claim that non-constative uses are in some way derivative from constative uses.

If this claim is true, then there is some reason for regarding the constative uses as the basic uses of mental ascriptions and hence some plausibility in taking the Sellarsian, or theoreticalist, account seriously, since the uses of theoretical ascriptions are constative, as I have defined this term. If, however, the claim is false, then there is some reason for resisting the theoreticalist account in so far as it does not capture, nor allow for, the non-constative uses of mental language. (This point will have to be argued later; I introduce it now simply to reveal my strategy in what follows.)

Care must be taken in the explication of the claim that non-constative uses are derivative from constative ones. I
shall divide the possible interpretations of the claim into two broad categories: the logical and the pragmatic. The logical interpretations involve the suppositions that non-constative utterances entail, imply or presuppose the proposition asserted by the corresponding constative use. It might be argued, for example, that if I use the expression "I wish you would go away" as a way of enjoining someone to depart, my utterance implies that I do wish her or him to go away; indeed, it might further be said, it is her or his recognition of the fact that such is my desire which is operative in bringing her or him to obey the injunction, if this s/he does, or to ignore it, if this s/he does.

My objection to this way of construing the situation is simply that logical relations hold between propositions, or statements expressing propositions, or even sentences used to make statements, but not between injunctions and propositions, nor between questions and propositions. The appearance of an implication in the present example is deceptive, for we may note that the same features occur in the example of the simple command "Go away!". In some way, the use of this imperative "gets across" the idea that I wish him/her to whom it is directed to depart, and if s/he goes it will probably be by recognition of this fact. But few would suppose that "Go away!" entails, implies or presupposes the proposition that I
wish someone (him or her) to go away.

I would prefer to say, using Austinian terminology [41], that it is a condition of the felicity of both the injunction and the command that I do in fact wish the relevant person to go away. One of the ways in which such an injunction may misfire is if this condition is not fulfilled. Of course, the injunction may still be effective in getting her/him to go away (against my wishes), but there is something amiss with the injunction - it is insincere or whatever.

Felicity-conditions are not truth-conditions, which is why entailment is irrelevant, but the same point may be expressed in terms of truth-conditions: one might say that it is a condition of my sincerely enjoining you to go away, of my sincerely using the expression "I wish you would go away" in this fashion, that I do wish you to go away.

The reader may be left with the impression that it is perfectly alright to speak of felicity-conditions being implied by non-constative uses of language such as those above. This is permissible, provided it is remembered that such an implication is not an entailment between what is said and the fact which constitutes the fulfillment of the felicity-condition. The proposition which asserts the existence of this fact is not part of the meaning of what is said, since the latter is non-propositional and no proposition
can be derived from it. For the same reason, the felicity-conditions are not presupposed by a non-constative use in any direct sense.

The sincere use of a non-constative utterance implies its felicity-conditions in the sense that, confronted by such a use, it is reasonable for the hearer to assume that those conditions are fulfilled, ceteris paribus. This is because in making such an utterance in your hearing, I let it be known that I make such an utterance and let it be assumed that I make the utterance sincerely. In that case you are entitled to assume, in the absence of indications to the contrary, that the relevant felicity-conditions are fulfilled. This is an inference not from what I say, but from the fact that I say it.

The claim that non-constative uses of an expression are derivative from its constative uses (where these exist) could be formulated in more pragmatic fashion, as the claim that unless the constative use were possible, unless, that is, it is a recognised move in the language, the non-constative use would not be possible. Upon close examination, however, this claim has little to recommend it. I see no reason why the form of words "I wish you would go away" should be any less effective, as an utterance with injunctive force, in the absence of the possibility of the corresponding use than in
its presence. Again the comparison with grammatical imperatives is useful, for "Go away!" does not suffer by the absence of a constative use. The pragmatic version of the derivation-claim is in many cases motivated by the idea, already mentioned, that it is by the recognition of some fact (expressible by the relevant constative use) about the speaker (in this case) that the non-constative use is successful. However, this fact can be recognised without our requiring the same form of words to express it. Furthermore, there seems to be no reason why the import of the non-constative use could not be picked up without the recognition of some such fact.²

If it be granted that there are non-constative uses of mental language independent of, yet existing side-by-side with fact-stating uses, the next task is to investigate the significance of this state-of-affairs for the alleged theoreticality of mental terms. We should observe, first of all, that the global feature of mental language³ to which I have been alluding, namely its employment in non-constative uses, is by no means generally found in other areas of theoretical discourse.⁴ Here, then, is some reason for refusing to regard mental terms as theoretical.

I have been stressing a disanalogy between the uses to which mental language can be put and those to which language which is non-controversially theoretical can be put. Those of
a theoreticalist frame of mind may question this procedure, for they may believe that all language is theoretical and that therefore theoretical language (viz. that part of it which is employed in everyday non-scientific discourse) does indeed exhibit the non-constative features which I have been using to drive a wedge between theoretical explanation and mental discourse. While I think this belief is false in the general case in the same way as in the mental case - it emphasises surface-similarities at the cost of ignoring deeper differences - I cannot hope to convince everyone of this by blank assertion, nor is there space to argue the matter out. I shall simply assume that the theoretical/non-theoretical distinction has real application to uses of language; I do not mind if this distinction marks a difference only in degree, for it is no less a distinction for that.

Having outlined the non-constative uses of mental language, I now turn to the question of whether these uses, as well as the function of mental language in reason-explanation described earlier, can be accounted for on the supposition that mental language is theoretical, in the sense explicated by Theses II and III. The emphasis initially is upon Thesis II.
In Chapter 4 and section (a) of this chapter I have canvassed two kinds of function for mental language, neither of which consists in the theoretico-causal explanation of behaviour, although one of them, the first, is often called "the explanation of behaviour". The existence of these two uses of mental language seems to me to constitute sufficient grounds for denying that the theoretico-causal model is adequate as a characterisation of the function of the mental, but this conclusion will be too quick for some. What I feel is the most plausible (and most common) justification for resistance to my conclusion is the view that both the systematic role of mental language in non-constative utterances and its function in the justification of actions are accretions parasitic upon its fundamental function, namely the causal explanation of behaviour. I now turn to an assessment of this possibility.

We should note first that the existence of the excrescent functions marks a strong disanalogy between the use of mental language and other areas of language which are admitted to be theoretical. We simply don't find such parasites there. Of course, while this is a good reason for
thinking that mental language is somehow special, it is not a conclusive reason for refusing to regard it as theoretical.

A more promising line of approach runs through the question of whether the so-called parasitical functions could exist without the so-called host function. I argued in section (a) of this chapter that this was possible in the case of the non-constative uses: although the non-constative use of "X has M" (for some subject X and mental-state-term "M") presupposes, in general, the fact that X has M, it does not presuppose the fact that these words have the function of asserting the existence of the fact that X has M. Thus the non-constative use could survive the death of the constative use. The same is true, I now wish to argue, in the case of the use of mental language in the justification of action. The citing of the subject's mental state as the (internalised) reason for his/her action presupposes the fact that the mental state was causally relevant to the action, but not the fact that there is a use of the relevant mental expression in the causal explanation of behaviour. Again the justificatory function could outlive the causal/explanatory function. Thus while the former function may be historically a parasite, it is logically independently viable.

This is an important conclusion, for it represents the different functions of mental language as conceptually
independent (in a certain sense) and hence increases the difficulty of any eliminative materialist who proposes to do away with such language. What the conclusion suggests is that the battle for the life of mental language must be fought not only in the arena of explanatory value and predictive power but also on the field of general human interaction, moral and otherwise, where exist the additional functions of expression, justification and so forth. This point will be developed later.

Another argument in favour of the conclusion that the present use of mental language is not theoretical can be constructed out of my analysis of ordinary-language explanation. This I represented as the rendering rational of the agent's actions. If one wishes to generate from this a full-blown causal explanation of behaviour, as opposed to the more or less informal assignment of "a" or "the" cause of an action to a mental state, one must make an assumption about the agent, that s/he is rational. D.C. Dennett [42] has argued persuasively to this conclusion, although his notion of rationality is more restricted than mine. As he points out, the assumption of rationality in the case of any particular agent may be amended in the light of further knowledge about his/her quirks and fancies, but this fact does not destroy the general point. The question I wish to ask is "How can the
assumption of rationality be included in the picture of mental explanation as theoretico-causal?"

It seems to me implausible to regard the rationality of the agent simply as what I shall call a base-condition in the causation of behaviour, as for instance a gun's being loaded is a causal base-condition of its firing when the trigger is pulled. There are circumstances in which it would make sense to say that the cause of the gun's going off was its being loaded, but none where it would make sense to say that the cause of a person's action was her/his rationality, in the minimal sense of rationality here involved. The grounds on which we ascribe rationality to an agent are that, in general, her/his actions are appropriate to her/his environment (beliefs) and goals (desires). It is thus not ascribable independently of the agent's actions and mental states (meaning by this last phrase beliefs and desires in the extended senses of these terms).

This fact is not itself damaging to the status of rationality as a causal condition. Although the rationality of the agent cannot be identified independently of her/his mental states and actions, it can be posited on the basis of past behaviour, and thus construed as a theoretical state for the explanation of her/his present behaviour. The problem is, however, that rationality, even in this minimal sense which
involves no evaluation and the most meagre portion of normativity, is a functional concept, not a structural one. It is not simply a causal condition under which beliefs and desires cause actions, but the appropriateness of those actions to those mental states.5

At first sight, then, the assessment of actions as rational requires the ascription to the agent of mental states. Thus rationality and the particular mental states of the agent are not logically independent theoretical constructs, if they are such at all. This of course does not show the theoreticalist account to be wrong - it merely shows that the activities of assessing actions as rational and ascribing mental states stand or fall together.

The appropriateness of action to mental state which is implied in rationality is linguistically given. We did not have to discover which actions are appropriate to which mental states, this being written into the very description of the mental state (as the desire for X, the belief that A-ing will produce X and so on) and sometimes of the action.

A further argument for the non-theoretical nature of mental explanation branches from this point. Attention to the linguistic situation enables us to explain why actions cannot be explained by inappropriate mental states. In the absence of some further beliefs (or facts) about the relation between
A-ing and B-ing (which, let us suppose, are extremely disparate), the desire for X and the belief that A-ing will produce X cannot explain my B-ing, since the mental states offer no rationale for the action. On the theoreticalist account, however, it would surely be a possibility, to be discounted by experience, that the above states cause the above action, no matter how inappropriate it might be. But that in normal circumstances the belief that drinking quenches thirst does not cause one to eat sand is not a fact which required discovery. It is a truth which is written into our ways of identifying and individuating mental states. Hence this way of elaborating the theoreticalist account is wrong.6

The argument of this section, then, runs as follows:
(i) the non-constative and justificatory functions of mental language are independent of (could exist without) the causal-explanatory function; (ii) in so far as mental ascriptions are used in causal explanation they require the attribution to the agent of rationality - otherwise they just won't work; (iii) rationality itself is not a hypothetical construct - it is the appropriateness of actions to mental states; (iv) what enables us to offer causal explanations of actions in terms of mental states is the fact that ascriptions of mental states may be used to justify actions; hence (v) if either use of mental language is parasitic upon the other, it is the causal role
which is parasitic upon the justificatory one; furthermore (vi) our ways of individuating mental states (with reference to the actions they justify) rule out the interpretation of these states as hypothetical constructs. It seems to me that it follows from (v) that Thesis II is false, and from (vi) that Thesis III is false.

It might be objected that here I base my conclusions on the role of mental ascriptions in accounting causally for actions, not for movements, it being the latter which are relevant to the idea that ordinary mental language is a theoretical competitor with neurophysiology. My reply is that causal explanations of movements in terms of mental states operate via actions, as argued in Chapter 4, so that (v) above remains true since (iv) is true also of causal explanations of movements. Given this, we have even more reason to oppose the view that neurophysiology and ordinary mental language are competitors, since the most important function of mental ascriptions of this type is the justification of actions, not the causal explanation of movements. Hence Thesis IV is false.

If the above argument is correct, Theses II-IV are all false, and so therefore is Thesis I. Even if it is incorrect, or if Rorty were willing to abandon Theses I-IV as part of his view, something of value emerges from the discussion of (i) above. In the next section I shall suggest that the existence
of mental states appears to follow from the validity of the roles of mental language mentioned in (i), and argue that these roles could not easily be taken over by neural language.
(c) Non-theoretical Uses and Neural Talk

As I argued at the end of section (a), the successful employment of a non-constative mental utterance implies the existence of a corresponding mental fact. For instance, if "I hope he'll come" is truly an expression of my mental state, then I must in fact so hope, even though the purpose of my utterance is not to report this fact. Similarly, that some mental state of mine is truly an (internalised) reason for my action implies that it has a role in the causation of that action, even though the giving of a reason does not have the purpose of reporting this fact. Even though the non-constative and rationalising uses could exist without the constative and causal/explanatory uses, it appears to be a condition of their successful employment that there are mental facts, and therefore mental states, which are causally efficacious. This provides us with a prima facie argument for the existence of the mental, granted the validity of the non-constative and rationalising uses.

A number of reactions are possible from the standpoint of Rorty. First he could suppose that ordinary language will be retained for the non-constative and rationalising roles while neurophysiology will become the language of reporting
and (causal) explanation. This alternative is not very attractive, since it does not question the validity of the \textit{prima facie} argument, in which case the existence of the mental remains intact. Our silence about the mental in reporting and explaining would not be sufficient grounds for denying its existence. The interesting reactions are those which involve challenging the \textit{prima facie} implication.

A way to do this consistent with the first reaction would be to graft the felicity-implications onto neurophysiological states instead of mental ones. Thus, for example, it would be a condition of the felicity of my expression of pain that my C-fibres were stimulated, or whatever. While I admit that such a change in the use of language is possible, I see two problems with this approach. The first is that, for Rorty, the regrafting must take place in a way which avoids the conclusion that the contingent identity theory is correct, in other words we cannot suppose, in remaining consistent with this version of eliminative materialism,\footnote{7} that a one-one correlation is possible between certain neurophysiological facts and the mental facts formerly included in the felicity-conditions for the non-constative utterances and the truth-conditions of the rationalising utterances. It is not clear to me that this is at all possible - certainly it would require a wholesale revision of
the non-constative and rationalising uses, and this may not be a price we wish to pay. The second problem is that such a regrafting would not be practical, since the new implication-relations would have to be learned piecemeal, there being no obvious connection of meaning between the antecedents and the consequents - the same substantive term never occurs in both, for instance. This impracticality does not exist with present mental language.

The second type of reaction involves the supposition that sentences in the language of neurophysiology could take on the non-constative and non-causal functions of mental language, so that now this would be replaced all the way across the board. If this were possible, it would reflect back on my claim that it is on account of the so-called "secondary" functions that mental language is not theoretical, since neurophysiological language would have those functions and it is theoretical, or so it appears. This conclusion is, however, harmless, since all such an eventuality would show is that neurophysiological language, once it has acquired the additional functions, will cease to be (merely) theoretical.

I do not think the linguistic change here considered is possible in any real sense: the alternative suffers the same problems of impracticality as the previous one - it simply is not clear, for example, that a certain neural formation is a
justification for slamming the door in someone's face. In other words, the logic of neural justification is not perspicacious (to us now, and as we imagine we will be) unless we fill in the gaps by talking about beliefs, desires, intentions etc.. Further problems arise from the open-ended nature of rationality; because it does not consider such factors relevant, eliminative materialism, in this version, gives us no picture of what moral dispute would look like, of what sorts of consideration could be urged in favour of each of a number of competing views, of why they should be regarded as relevant. Our present systems of justification are admittedly frayed at the edges, but that is as it should be and most people have no difficulty operating with them. What we are faced with if we wish to embrace this version of materialism is a significantly different system which in our present state of knowledge is mud-clear to us. We do not know what we are being asked to embrace, nor, therefore, if it is possible to do so.

I say that whatever systems of justification are to replace our present ones must be significantly different because if this possibility is to be taken as supporting the relevant version of eliminative materialism we are constrained to avoid admitting the truth of the identity theory, as in the case of the previous reaction. The replacement of mental
language with neurophysiological language will be a Pyrrhic victory for Rorty's view unless there are significant differences between the two uses - otherwise mental states will have survived to be called by different names. It is to be expected that these differences will extend beyond the causal/theoretical arena of discourse into that of justificatory explanation. While none of the above considerations shows that the much-vaunted replacement is utterly impossible, they widen the grounds upon which such a linguistic change is to be considered. Significant changes in our systems of justification are significant moral (etc.) issues, and they are not to be decided on grounds of theoretical parsimony, predictive value and so forth, which would be the sorts of criteria relevant if the issue were a purely scientific one, as Rorty appears to claim.

Both the reactions treated so far accept the validity of the non-constative and justificatory uses of mental language but ignore the claimed implications of these for the existence of mental states - these they replace by implications for the existence of neural states. It seems to me that these projects are hopeless, although I have not proved this, because the relevant implications for the existence of mental states are essential to the non-causal uses. A third reaction is available to the eliminative
materialist, and this I believe is for him/her the most consistent, and that is to admit the implications but to reject the non-constative and justificatory uses of mental language along with mental states. I see no conceptual problem with this alternative, but I would insist that the abandoning of such linguistic functions is a radical moral move and that moral considerations are relevant to its advisability. (I concentrate upon moral justification because it is arguably the most important, but what I say applies to aesthetics and much else besides.) This general topic will be revisited in the concluding chapter.
Footnotes to Chapter 6

1. This qualification is inserted to assuage the objections of eliminative materialists who might argue that its omission renders it false, in the light of the non-existence of mental states, that any such sentence can state a fact. It will clearly be tedious to continue to make the qualification, so I shall omit it in concreto while admitting it in abstracto.

2. Two opposed views of language confront each other here. The first regards constative utterances, and the activity of stating facts, describing the world, as basic, as the main function of language. The second view allows an independent function to non-constative utterances. There is even some plausibility, it seems to me, in the view that it is the non-constative utterances which are basic, if anything is. The view pays great attention to the role of language in human interaction. However, I cannot argue out this issue here.

3. It is relevant that the feature to which I am drawing attention is global. That mental terms occur in non-constative utterances is not by itself significant - the existence of the imperative "Eat the lobster" has no implications for the special status of lobsters. What is significant is the systematic way in which many non-constative utterances include a mental term as part of the main verb-phrase, as in the examples cited.

4. The only example of a theoretico-explanatory context which I can think of where the terms of the discourse can be (and are) used to make non-constative utterances is psychiatry, for instance psycho-analysis. It is characteristic of such forms of explanation that those who engage in them, whether as therapist or patient, quickly acquire the ability to express, enjoin, warn etc. in the language of the theory as opposed to that of everyday life. However, since the differences between these two areas of language are so small, this "opposition" tends to evaporate into a complementation. As a result of this, and of the fact that in going over to psychiatry we have not moved away from the mind, the existence of psycho-analytical (and other) explanation, even if this is straightforwardly causal (which can also be questioned), does little to help us avoid the general disanalogy, with respect to the possibility of non-constative uses, between mental language on the one hand and the language of other theoretical disciplines, such as the special sciences, on the other.
5. This point is (implicitly?) acknowledged by Dennett [42], who, while holding an eliminative materialist view of ordinary mental concepts, believes they will (ought to?) be replaced not by those of neurophysiology, but by the functional concepts of machine-theory; this would leave room for the application of a concept similar to means-end rationality, but not, I should argue, for the more evaluative notion of moral rationality.

6. This conclusion may not be seriously damaging to Rorty's eliminative materialist position. The claim that the present use of mental language is theoretico-causal can be seen as an inessential adjunct to that position, one move in the attack on ordinary-language explanation. That this involves a broader conceptual weave might simply be taken to mean that things are so much the worse for it, might provide us with more grounds for replacing it with a more scientific discourse. The important point for my side of the argument is that if the claim above is false, the issue between ordinary-language and neurophysiology is not one that can be settled by appeal to the canons of scientific methodology alone. It is not a struggle between two scientific theories, but one between science and ordinary life for the determining role in ontology. In the final chapter I shall suggest a peaceful solution to this problem.

7. This point is argued in Chapter 2, section (d).
(a) Summary

My aim in this essay has been to refute the version of eliminative materialism which in Chapter 1 I attribute to Richard Rorty. I begin in Chapter 2 with the explication of his view, contrasting it with what I take to be the most sophisticated form of the (neural) contingent identity theory, due to Lewis. According to Lewis, certain parts of the physical will come to be identified with the mental by being discovered to occupy the same causal role. I believe that this version of the identity theory is relatively safe against all standard objections.

It is not, however, a successful view, for it falls to what I call the "Surroundings" objection, drawn from some remarks of Wittgenstein and Malcolm with some help from the work of Tyler Burge. What this objection shows is that the truth-conditions for mental ascriptions spread out much farther than the confines of the central nervous system, or indeed than the collection of non-relational facts about the subject her/himself, with the result that, given what we already know about the present use of mental language, it will not be possible to formulate true correlation-statements of
the kind envisaged to provide the observational base for the neural version of the identity theory, which is thus refuted.

Physicalism does not however die with its historically most popular embodiment. There are three main alternatives: (i) one could say that the social and environmental facts relevant to the truth-conditions for mental ascriptions could themselves be reduced to physical description and take their place alongside the neurophysiological facts on one side of the equations of a new identity theory. Such a theory should be immune from counter-examples (at least of the sort which killed its predecessor), though it might be attacked on grounds which undermine the significance of insisting on identity where manifestation, embodiment or mere correlation will do. What seems more plausible to me is (ii) eliminative materialism. It is no objection to this view that the relevant correlation-statements cannot be found: rather the opposite, for it is the absence of such correlations which provides a reason for discarding mental concepts in the face of the scientific superiority of neurophysiology. The other alternative is (iii) a form of non-reductive materialism such as that espoused by Hellman and Thompson [43] [44], the discussion of which is outside the scope of this essay.

Eliminative materialism is thus shown to be a natural reaction to the demise of the identity theory. In the
remainder of Chapter 2 I set up the version of the view which I intend to attack, and make some enquiries about a suitable methodology for the attack. The main representative of this view is Richard Rorty (whose more recent writings [37] can be interpreted as indicating that he has abandoned the view), and it is upon his exposition that I concentrate. I suggest that what Rorty requires to support his view are (i) an incompatibility between mental and physical and (ii) an incompatibility between ordinary mental language and that of neurophysiology, and that he grounds the second in the first, and this in the incorrigibility of certain mental reports. I argue that there is no reason to expect the incompatibility mentioned in (i) - that Rorty's argument for it is unsound - but that it is essential to his view. My characterisation of his position is achieved largely through a consideration and an elaboration of the demon analogy.

As part of this attempt I try to clear up some apparent inconsistencies in Rorty's view regarding the conditions under which it would be rational to give up "mind-theory" in favour of neurophysiology. I argue that these include (i) the claim that mental concepts involve a commitment to incorrigibility and (ii) the failure of perfect psychophysical correlations. I then outline the strategy of my attack.

Chapter 3 is devoted to an assessment of the claims
that incorrigibility is the mark of the mental and that this grounds the incompatibility required for the elimination of mental states. After showing why the non-existence version of this view is superior to the disappearance version, I argue that (i) mental ascriptions (of the relevant type) are not incorrigible in Rorty's sense (nor in the traditional sense), (ii) Rorty's form of incorrigibility fails to distinguish between mental and physical, in that it applies to some physical ascriptions, (iii) even if Rorty's form of incorrigibility were the mark of the mental, it could not ground the categorial incompatibility which he seems to require, and therefore that (iv) Thesis VII is false.

In the final part of the chapter I offer a summary of the criterial view of mental discourse which I have extracted from Wittgenstein and explain how this accounts for the form of incorrigibility which I think is characteristic of certain mental reports. This account is necessary for some of my arguments in Chapter 5.

The next three chapters, 4, 5 and 6, are devoted to an assessment of the extent to which Rorty's view can account for the various roles of mental language which I consider worth mentioning. These include the explanatory and reporting roles deemed relevant by Rorty, but also reason-explanation and the broad class of non-constative utterances. My general position
is that the claimed theoreticality of mental states (as expressed in Theses I-IV) renders his view incapable of offering an adequate account.

The claim that mental terms are theoretical terms I analyse into two components, which I treat separately. They are (i) that the (primary) use of mental terms is in the causal explanation of behaviour, and (ii) that mental states are hypothetical entities, in the sense that as far as we know they may not exist.

Chapter 4 examines the theoreticalist view of mental language (which Rorty adopts from Sellars) with the purpose of establishing that the view is plausible as a view about the causal explanation of physical movements, though not of actions. I then contrast the causal explanation of movements with reason-explanation, which consists primarily in the explanation of actions, arguing that it is an independent role for mental language. I also suggest that some features of explanation in mental terms make it plausible that it is not incompatible with neurophysiological explanation, even in so far as it is causal. Here I refer back to some material from Chapter 2 (where the same point was made as part of the elucidation of what these Rorty must deny to establish his own view).

It is important to be clear on the manner in which this
compatibility is achieved: it is not that both mental and neural facts occur in the same theoretico-causal explanations of behaviour, but that the behaviour can be successfully explained using either set of terms. We have a difference in levels of explanation, not an integration of explanatory nets. I explicate the relationship between mental and neural systems of explanation by reference to the analogy between this case and that of the existence of both hardware and software levels of the explanation of the behaviour of complex machines.¹

Chapter 5 turns to the reporting role of mental language. My goal is to show that Rorty's explication of mental reports is inadequate. Before this I argue that it is a mistake to ground the existence of the mental in the phenomenal immediacy of sensation, i.e. on the Myth of the Given - thus I eschew what I call phenomenism, and with it idealism, phenomenalism and the Cartesian picture of the mind. It is essential for the reader to realise that although I am in some sense a mentalist, I am not a phenomenist etc.. If the existence of mental states is given, it is given in the use of mental language, not in how things seem to me. I stress this aspect of my view because it has usually been assumed by materialists that some form of phenomenism is the only alternative to their view, and it is important for me to dispel this illusion lest I be misunderstood.
Even when one abandons the Myth of the Given, however, one is left with the strong intuition that when, for example, I attribute a headache to someone, I sometimes speak truly. This intuition seems to be denied by any literal reading of Rorty's version of eliminative materialism. Surprisingly, we find when we look at the writings of various of that school that those philosophers exert great effort to accommodate their views to that intuition. In view of the failure of that exertion, a failure which I believe I have demonstrated, it would perhaps have been best for Rorty to chew the gristle with the lean meat by denying the truth of ordinary mental ascriptions. His view might then be less popular, but it would surely be no more counter-intuitive than it already is on account of its denial of the existence of mental states.  

The argument of Chapter 5 is an attack on the conciliatory efforts of Rorty's eliminative materialism - it would have no application if these were not made.

What Rorty needs, then, is a way in which ordinary mental ascriptions (reports, in his terminology) can be true even though mental states do not exist. Two possible moves, which I mention and discard, are the appeal to a pragmatic theory of truth (which I have heard Rorty make in conversation) and Savitt's claim that existence-statements are ambiguous, depending on whether we take them as "internal" or
"external" to some theory. What I believe to be the more plausible attempts, and the ones which Rorty's published writings appear to underscore, involve the notion that mental reports are really about neural states, and so can be true in virtue of the neural state of the subject even if mental states do not exist. Hence the bulk of the section is spent in arguing against the formulation due to Shope (in terms of the distinction between "report of" and "report on") and the more general causal theory of reporting. I argue that neither of these analyses can cope successfully with false mental reports, and are therefore inadequate as analyses. I give some further reasons for thinking they will not do even construed as ad hoc attempts to save certain "true" mental reports.

A further argument against this theoreticalist account of reporting is introduced with the notion of incorrigibility. While I reject the phenomenist conception of incorrigibility, and so do not criticise the present version of eliminative materialism for failing to live up to that conception, I am no more enthralled by the theoreticalist substitute. The notion of "empirical" incorrigibility does not seem to me to fit the facts which give rise to the idea of incorrigibility - it allots too much epistemic priority to first-person present-tense reports. I also argue at great length that Rorty's
published account of the genesis of (his form of) incorrigibility is seriously incoherent, and consider and reject various possible ways of making up this deficiency. The main point here is that any view which attempts to justify giving epistemic priority to a form of evidence (first-person reports) which is merely the causal result of a more fundamental evidential connection (between mental ascriptions and overt behaviour) must be mistaken, for the more fundamental connection will always deserve priority in cases of conflict.

Now in other places Rorty flirts with what we both call the "Wittgensteinian" notion of incorrigibility, he not distinguishing it, apparently, from the empirical form. I try to show how the two notions are distinct, the former denying, while the latter requires, a difference, detectable at this stage of conceptual development, between being linguistically competent in the making of a mental report of the relevant type and speaking truly (the sincerity of the report being granted). The Wittgensteinian notion, which I believe fits the facts of language use best, differs from the standard phenomenist conception of "logical" incorrigibility in so far as it allows that we might come correctly to assert the existence of this difference through neurophysiological discovery and the conceptual change which might thence ensue.
I argue that Rorty's "empirical" form of incorrigibility presupposes that this difference exists and is detectable without appeal to neural evidence, but rather on the basis of standard behavioural evidence. This feature is disguised by Rorty's choice of examples, but is logically required by his view.

Having thus established a difference between the "empirical" notion and the "Wittgensteinian" one, I argue that the latter is incompatible with some of the broader features of the theoreticalist account, and so cannot be appealed to as an alternative to the already-rejected treatment in terms of "empirical" incorrigibility. Thus Rorty's accounts of incorrigibility, and of mental reporting in general, prove to be seriously inadequate, and the attempts to rescue our ordinary intuitions about incorrigibility, as well as the attempts with regard to the truth of mental ascriptions, fail. He must abandon those intuitions to save his view, but obviously this merely increases its implausibility.

While Chapter 5 removes much support from this version of eliminative materialism, it does not supply a knockdown argument against it. This is not surprising, for, it would seem, no amount of logic can demonstrate that mental states exist (the same applies to trees!). The most that one can do is to defuse the eliminative materialist arguments in favour
of the conclusion that the denial of the existence of mental states is a rational reaction to the (non-correlational) supersession of mental talk by neurophysiology.

The two roles of mental language which Rorty does not consider are tackled in Chapter 6. I should like to have supported a position, held by many philosophers of the British ordinary-language school [45] [46], according to which the relation between mental states and behaviour is in no important sense causal, but I have found this position too implausible to sustain. However, while I do not in this way approach a knockdown argument against the theoreticalist account, I do reach some relevant conclusions which broaden the scope of the problematic in a manner which enables me to add further support for my own position.

It will be remembered that I examine two main types of use for mental language which do not consist in causal explanation. The first is the systematic way in which mental language is employed in non-constative utterances, notably those involving expression (as opposed to reporting). The second, already discussed in Chapter 4, is what we ordinarily call "mental explanation" - the giving of reasons for action. I argue that the latter activity, in so far as it involves mentioning the agent's mental states, is a mode of justification, not of causal explanation, and that both
"anomalous" uses of mental language are conceptually independent of the role of mental language in causal explanation. It would be possible for the anomalous uses to survive the abandonment (e.g. as a result of theory-replacement) of the causal use. It does not follow from this, however, that there is no relation between the anomalous uses and the causal facts. On the contrary, it appears to be a condition of the felicity of most forms of non-constative mental utterances that the relevant subject be in some mental state (which, arguably, plays a role in the causation of the relevant utterance), and a condition of the appropriateness (propriety, acceptability, correctness) of a justification in mental terms that the subject be in the mental state mentioned in the justification and that this be involved in the causation of the action to be explained.

This curious result supplies me with a very quick argument from the validity of the anomalous uses to the existence of mental states (though not to the continued use of mental language in causal explanation). (The result is, as it were, very consistent with my conception of the compatibility of causal explanation in mental and in neural terms.) At the end of the chapter I consider various possible eliminative materialist reactions to this argument, with more or less the same conclusion in each case: it must be supposed either that
neural language will take over the anomalous functions of mental language, or that these functions will be abandoned also. In the first case a certain amount of restructuring of our systems of justification will be required (since, _ex hypothesi_, there will not be a straightforward mapping of mental states onto neural states); in the second case the restructuring, being terminal, is even more pronounced! In both cases the replacement of mental language by neural language will involve significant moral moves, significant changes in what we count as a case of justified action.

The importance of this conclusion for my general project is that it shows that the question of whether, given the theoretical superiority of neural explanation, mental language ought to be replaced, is to be answered not only on scientifical-methodological grounds, but also on moral and ethical grounds. The price we have to pay for the replacement may be much higher than the "inconvenience" with which Rorty identifies it. This of course has not been _established_ - it may be that our moral intuitions _need_ the sort of reorganising that would occur as the result of the widescale adoption of neural talk, so that we should be glad to make the move. My own intuitions run counter to this, but all I can say against Rorty is that he has not considered in anything like enough detail the conceptual consequences of the replacement of
mental language, and that until he does so we are not in a position to assess his view with any hope of completeness. This is a programmatic objection.

Before making it I offer an argument directly against Theses II-IV, on the grounds that the relationship between the justificatory and causal-explanatory uses of mental language is such that the former is primary, in the sense that the concept of rationality which it underwrites is necessary for the success of the latter. It would be advisable here for me to explain my view of the causal role of mental language.

It seems to me unobjectionable, and perhaps therefore somewhat trivial, to allow that mental states stand in causal relations. While I am not prepared to lay out here a theory of causality, I should admit that I operate with a fairly wide concept of cause, which is such as to render incontestable the causal efficacy of mental states; this efficacy, however, is as insignificant as it is incontestable. Allow me to elaborate: I have said, for example, that it is a condition of the correctness of a reason-explanation that some corresponding causal statement be true - if "she did it because she believed him a villain" justifies her turning him out of doors, it must be true that her belief that he was a villain figured causally in the production of the action - but this is just to say that the correctness of the reason-
explanation licenses another form of statement, which I am willing to call "causal". This willingness is informed, however, by caution about treating mental causes of this type importantly similar to, say, causes within the ken of physics, for there are obvious differences. For instance, there is a serious question whether backing up a reason-explanation by the corresponding causal statement adds anything at all to the original explanation. This, I believe, is connected to the point that it would be wrong to suppose, of most mental explanations, that they were discovered to be (theoretically acceptable) explanations. (One can discover that they apply to a particular case, but one cannot have discovered that, in general, the desire to kill causes one to kill.) Another point is that mental explanation allows for the existence of other parallel levels of causal explanation, such as neurophysiology: this feature also exists in the natural sciences, though arguably less prominently; I mention it here because dangers arise when one assumes that mental explanation is an attempt to give the causes of a piece of behaviour, implying uniqueness.

Unless one preserves this caution about assimilating causes, and levels of causal explanation, of different types, one can be led into error. One may find oneself believing that mental states simply are not causes, or that mental
explanation is inadequate as a form of causal explanation, because, for instance, it does not yield rigorous prediction of the kind highly-prized in the natural sciences. Both of these (opposing) conclusions are based upon too strict a notion of cause, and both are wrong.

If this conception of causality be granted, I feel I am entitled to conclude that Theses I-IV are false. The matter is, however, rather contentious, and I may have to retreat to the programmatic objection outlined above.
(b) Conclusions

(i) Rorty's View Refuted

Brands of eliminative materialism may be divided across two parameters. One is the grounds upon which the view is based: the version treated here has as one of its fundamental assumptions the claim that there is an incompatibility between mental and physical. The other is the attitude taken towards the eventual elimination of mental talk. It is felt that this is inevitable, desirable or merely possible. Different treatments are required, according to which attitude the eliminative materialist wishes to support. Each of these attitudes is supportable by the views which, by the first criterion, fall within the scope of this essay, and so I shall discuss each of them in turn.

So long as we are talking merely about the elimination of mental talk, and not about that of mental states, it is easy to agree that such elimination is possible. There are lots of things without talking about which human beings could get along, and there seems to be little special about the mental in this regard. If one is pessimistic about the chances of a continuation of the flourishing of the human spirit in the face of scientific discovery, one can also agree
that such elimination is inevitable. I do not agree, but this is a matter for empirical prediction, not philosophical argument. What speaks in favour of the inevitability is the success with which science has already revamped certain parts of our vocabulary; what speaks against it in the case of mental language is the fact that neural talk is likely to be much more clumsy to operate with (if the analogy with machine-explanation has anything to recommend it), so that mental talk may survive, if only as an heuristic device. It is of course admitted by Rorty that the replacement of mental language may involve a certain amount of inconvenience, so that this point does not speak against its possibility. There must be more to the view, though, than a possibility which no one disputes and an inevitability which is admitted to be empirically highly contentious.

Part of the story concerns whether it would be desirable to eliminate mental talk, and here I think that most eliminative materialists, Rorty included, tend to think that it would be desirable, except perhaps for the inconvenience involved. I oppose this attitude. I think these philosophers have been overly impressed by the goals of science (prediction, control etc.) and wrongly assumed that we should make these an important part of our goals in dealing with one another. More importantly, I think that the cost of the
elimination of mental talk has been underestimated when it is described as "some inconvenience". As the argument of Chapter 6 showed, if it is to occur in eliminative fashion, as opposed to via the identification of mental states with neural states, the replacement will require wholesale changes in our systems of justification, moral and otherwise, which it is not clear we should be willing to accept. I am in fact convinced that the elimination of mental talk is not desirable - the differences between eliminative materialists and myself here centre on our pictures of how we should like human beings to develop. However, while there is much room for speculation on these matters, there is little for hard philosophical application.

To approach the difficult issues we have to discuss not only the possibility of the elimination of mental talk but also that of the elimination of mental entities. The real possibility that Rorty is backing is the possibility that it would be correct to deny the existence of mental states once neurophysiology has shown that an alternative mode of explanation for behaviour is possible. Again, no one would dispute the logical possibility of the falsity of "Mental states exist" - but this would be only to affirm the contingency of that statement. The question that we have to answer is this: "Given the ways in which human beings behave,
and the circumstances in which they find themselves, is it possible that mental states do not exist?" Of course I believe that the answer to this question is "No".

I have supported this belief by disposing of Rorty's eliminative materialist argument in favour of a positive answer. It should be remembered that the version of eliminative materialism I have attacked bases its answer on a certain thesis about mental concepts, the claim that mental concepts are irredeemably Cartesian, in the sense of involving incorrigibility. This claim has both a strong and a weak form. On the strong form it is a matter of definition that mental terms have Cartesian implications; I believe that this is easily refuted, but it is improbable that it is held by Rorty. The weak version is more viable: on this, the conceptual connection between the Cartesian and the mental is a result of its being more convenient, in the face of the eventual overthrowing of incorrigibility, to do away with the mental than to attempt to retain it by cutting away the now embarrassing implications. This weaker claim is the result of taking seriously the tradition, stemming from Quine, according to which it is not in general possible to distinguish between those features of a thing which it has in virtue of the meaning of the term used to refer to it and those which it is merely commonly believed to have.
The usual conclusion drawn from these considerations is that it is pointless to talk about meaning and that there is no real distinction between analytic and synthetic. I myself would dispute this conclusion, but what it is important for me to stress is that it cannot either be embraced by Rorty. For if all features of the mental are merely synthetically related to mental states themselves, then so too is their alleged incorrigibility. In this case the existence of the mental is not impugned by the possible discovery that nothing has this character.

What I do regard as a permissible conclusion from the Quinean premisses is that the distinction between features-in-virtue-of-meaning and features-in-virtue-of-belief (and hence that between analytic and synthetic) is a matter of degree. We place a feature on the meaning end of this continuum according to the degree to which we should be inclined to say that anything which lacks this feature fails to qualify under the relevant term. What is of interest here is not our individual personal inclination (save in certain atypical circumstances), but the inclination of our linguistic community (as enshrined in linguistic practice), which seems to be governed by linguistic convenience. Hence the weaker version of the incorrigibility-claim.

Rorty's view, when based upon this version, appears
quite plausible provided that we construe the parameters of linguistic convenience in terms of scientifical-theoretical superiority. On this eliminative view, what shows the "mind-theory" explanation of behaviour to be incorrect is the emergence of a superior theory, neurophysiology or something very like it. Therefore it seems that there is no need to retain mental states when the alleged commitments of mental concepts have been discredited, since the theoretical gap left by their absence can be immediately filled.

However, that there is no theoretical need for entities of a certain type does not imply that those entities do not exist, even if there is no possibility of identifying them with sets or configurations of theoretically-acceptable entities. What I deny here is a basic tenet of many forms of reductionism, that there is an ontologically-privileged class of entities (the Taken as Given? - usually the latest entities of microanalysis) such that anything else which exists is identifiable with collections of members of that class. It is the absence of identifiability in the case of mental states which leads Rorty, as eliminative materialist, to deny their existence. I suggest that this reaction is inappropriate, since it is not in general true that identifiability is a necessary condition for existence. The existence of tables does not wait upon (or even upon the presumption of) their
identifiability with clouds of molecules. People knew a thousand years ago that tables existed, and would not rightly have given up their knowledge had they known, per impossibile, that no such theory as the molecular theory of matter would ever come along to explain tabular phenomena in other terms (with or without identifiability).

So too, I claim, with mental states: their existence is much more certain than any account of what they are. In this matter, the contingent-identity-theorist has the advantage over Rorty, in so far as s/he has realised that in our attempt to discover what mental states are we are led by the phenomena, not by the apparent meaning of mental terms. The most plausible version of Rorty's view says, "If anything is a mental state it has certain properties F. It will soon be discovered that nothing has F; then we shall be correct in saying that there are no mental states." The identity-theorist says, "Ordinary ascriptions of mental states are very often true (and imply the existence of what is ascribed). If we can find out what goes on at the neural (etc.?) level whenever a particular ascription is true, we shall have discovered what that particular mental state really is." While I disagree with the latter's proposed account of what mental states are, I concur with her/him in allotting primacy to the phenomena.

Notice that I am not here relying on my belief that
mental states are not theoretical entities, in a broad sense. There are many examples within the realm of scientific theory of entities whose existence is now beyond doubt. Atoms are particularly prominent in this regard. What must be noted is that their existence is not certain because we have an account of the meaning of "atom" and proof that something qualifies by that account; it is certain because no future discovery will lead us to abandon the existence of atoms, though we may be led to change our account of the properties of atoms (even beyond recognition). I am inclined to think that this is a fact merely of linguistic convenience, but this does not prevent it from having ontological consequences.

The case against the eliminability of mental states, construed as theoretical entities, then depends upon how similar to the case of atoms is the present one. I, however, do not need to answer the question here implied, since I oppose the construal which it presupposes. On my "ordinary language" account of the meaning of mental terms the existence of the mental is guaranteed by the facts that (i) mental states are referred to by terms which can occur in sentences in subject position, (ii) mental language has successful uses — in particular people employ it to distinguish between cases they wish to treat differently, and (iii) no matter what future discoveries are made, none of them will be
revolutionary enough to force us to abandon the use of mental language. Here (ii) is a summary of the criterial account of mental language which I mentioned in Chapter 3, and which is discussed in the Appendix below; it should be noted, furthermore, that in (iii) I do not say that there is no conceivable discovery which might lead us to abandon mental states, only that no actual discovery will force our hand on this point. I do not think that the logical provisionality here accorded to mental states is any bar to our certainty about their existence, any more than it is in the case of statements about the physical world.

I thus presuppose an ontology of ordinary language. I do not deny that science makes important contributions to ontology, nor that the two cannot come into conflict. What seems to me obvious, however, is that most cases of conflict between the ontological designs of science and of ordinary language are merely apparent. The existence of the mental just is compatible with the existence of the neural, and indeed with the completeness of neural explanation. The eliminative materialist view under consideration opposes this conclusion by claiming (i) that there is only one correct account of the causation of behaviour, and (ii) that there is a conceptual incompatibility between mental and physical (which rules out the identity theory as a form of materialism
consistent with (i), yet affirming the existence of mental states). I believe I have shown that both of these claims are mistaken, and that I am entitled to conclude that the version of eliminative materialism under discussion has been refuted.

(ii) Rorty's View: the Grain of Truth

While the main conclusion of this essay is that Rorty's view, as a position in the philosophy of mind, is to be rejected, I wish to record my own sentiment that its supporters do have something worthwhile to say. This requires me to recast the view as one which is primarily concerned with epistemology. I have earlier remarked that the main object of attack for Rorty often seems to be not the existence of mental states but the Myth of the Given. I am sure that, at the time of his seminal writings, Rorty believed that mentality is just part and parcel of that Myth, but this is a belief I would challenge. What I propose to do here is to disentangle the epistemological and ontological concerns of eliminative materialism and indicate that a certain amount of success has accrued to the view with regard to those of the first kind.

The knotting together of these two kinds of concerns was first achieved by Descartes, and has become more or less
standard in phenomenist views. Here we have the
foundationalist conception of knowledge according to which our
knowledge of our own mental states forms the incorrigible
basis upon which all else is built. Rorty is right to attack
this conception, but wrong in thinking that in order for the
attack to be successful mental states must be impugned along
the way.

In the first place, the question whether knowledge is
foundational is distinct from the question whether the
(purported) basis is incorrigible. Hence it should be
attacked without bringing in the alleged incorrigible nature
of mental states. I see no reason why the existence of a
small class of epistemically certain statements should be
anything but a small obstacle to a general epistemological
holism, and this is even more clear when incorrigibility turns
out, on investigation, to fail to justify the kind of
categorical certainty that was hoped for. (See my arguments in
Chapters 3 and 5.)

Secondly, one can attack the incorrigibility-thesis
without impugning mental states. To prove that there is
nothing which is logically incorrigible is to prove that there
are no mental states only on the assumption that, as a matter
of definition (or something like it), mental states are
logically incorrigible. And this assumption is just not
justified, a point which by now will perhaps be too familiar to the reader.

Thirdly, even if one replaces the notion of logical incorrigibility with Rorty's watered-down "empirical" form, which more plausibly attaches itself to certain mental states, doing away with incorrigibility in this form is not sufficient for doing away with mental states. Certain religious societies may not question theological assertions made by their members, but the fact that at some time in the future these assertions will be questioned does not entail that the entities about which they are assertions do not exist.

In summary, then, I should like to say that I applaud the eliminative materialist strivings against the Myth of the Given - this is where the real work needs to be done, and where there is hope of success. The campaign against mental states is only an insignificant distraction, and one which is doomed to failure, from those endeavours.

(iii) Materialism: Where Next?

The main argument of my essay leaves open three possible positions to which a traditional materialist might retreat. I shall present these, indicate in what directions problems for them lie and recommend a more plausible position
which those of materialist persuasions might be willing to adopt, though it is not itself a form of materialism.

The first possible position is a brand of eliminative materialism. The versions of this which I have attacked share the common assumption that mentality is incompatible with physicality, and in particular that the existence of a complete neural explanation of behaviour is incompatible with that of one in mental terms. I have undermined this ground by appeal to the software/hardware analogy and by other means. To survive my arguments, an eliminative materialist must drop this assumption. S/he must allow that mental and neural explanations are compatible, but claim that the theoretical superiority of the latter is still sufficient grounds for showing that the entities appealed to in the former do not exist. That is, s/he must adopt a much more Feyerabendian position in the philosophy of science, suggesting, for instance, that the two modes of explanation are compatible because incommensurable. The problems with this approach are (i) that it is not clear that the falling into disuse of a theory implies that its theoretical constructs do not exist, and (ii) that the approach still relies on the conception of mental states as theoretical entities, a conception I have tried to dispel throughout this work.

The second possible position is a brand of the identity
theory which takes into consideration the arguments (against earlier versions) which I mentioned as "The 'Surroundings' Objection" in Chapter 2. The idea would be that mental facts are to be identified not simply with neural facts, but with collections of all those facts (or their physical counterparts) which enter into the truth-conditions for mental ascriptions. This would involve a significant particularisation of the proposed identities, to accommodate the relativisation to context and circumstances. It is not clear to me how far this process would have to be pursued; what is clear is that the farther it is pursued, and the more insulated from counter-examples the theory becomes, the less the significance of the identity-claims which emerge. There is always a certain physical situation concurrent with any mental state, but there is nothing of interest in identifying the two unless the identification carries over into other cases. Token-token identities can be constructed between any two contemporaneous events, but what are needed for a significant identity theory are type-type (or arguably type-token) identities (at least of some kind). It is not yet clear whether these will be available in the present case, and so I cannot yet pass judgment on it.

The third possibility is the adoption of a non-reductive brand of physicalism along the lines suggested by
Hellman and Thompson [43] [44]. I shall not discuss this view here, except to note that it does not seem to be essentially different from the rejection of materialism which my own view involves. It may, I suppose, be possible to find some difference which can breathe life into the now hoary opposition between materialism and mentalism, but what I should honestly recommend to the materialist is that s/he abandon materialism altogether. My main reason for this is that the demon which materialism fears has already been exorcised. I have noted elsewhere that it is an assumption of most materialists that the only alternative to their view is some form of phenomenism, involving the Cartesian conception of the mind as a non-physical (yet quasi-physical) mechanism. The issue, however, is not that simple. I reject the Cartesian model of the mind, but not because it involves the claim that the mind is immaterial: economies, jurisdictions, cultures, novels, dangers, promises, anxieties – all these and many more are immaterial things, that is to say, they are not material things, yet they are things. The Cartesian picture is to be rejected because, in a perverse way, it treats the mind as too like a material thing (the belief that minds can survive bodily death is one example of this). At the same time I do not find myself driven to embrace materialism, mainly because I oppose the species of reductionism which
often form a part of that view.

It is the reductionism which I am urging the materialist to abandon when I recommend that s/he give up his/her view. Everything else s/he can retain: the idea that it is "in principle" possible to give a complete causal explanation of the universe in physical terms, the idea that minds, in particular, do not "interfere" with physical causal processes (though they do stand in causal relations with physical events), the idea that nothing else would exist if physical things did not exist. What I am opposed to is basically the idea that immaterial facts either (i) don't exist, because physical facts can take their place, or (ii) are mappable onto complexes of physical facts. It is the open texture of our concepts employed to refer to immaterial things which prevents mappability, and if this is involved in reduction, then reductionism is false.

As I have noted, a brand of physicalism has arisen in recent years which does not presuppose mappability in any rigorous form. So long as this view allows for the existence of immaterial things and for the incommensurability of the relevant concepts with physical concepts, I have no quarrel with it. If it is to be called materialism, then I am a materialist. I wish to note only that it differs in important ways from traditional materialist/physicalist views.
(c) Speculations

My final task in this essay is to delineate some of its limitations and to indicate in what directions I should like to develop the anti-materialist side of the dialectic.

It is not uncommon for a philosopher to feel, at the time of completion of a piece of work, that it is already out of date. This feeling is perhaps most acute in the case of doctoral dissertations, which are written over a span of time during which the writer's ideas may be undergoing revision at a pace faster than that of the pen which records them. Something of this nature has been true of my own case. In order to avoid this constant shifting of ground, I found it necessary to take certain points as fixed, and to employ them as fulcra around which to turn the entire problematic. Thus I take myself not to have supplied the truth of the matter, but to have contributed to a stage in the dialectical process which constitutes the philosophy of mind.

What now seems to me most contentious about the underpinnings of the views here expressed is not the way in which my position differs from those I attack, so much as some of the assumptions common to all the standard views which I have myself adopted, for the sake of argument, as sacrosanct.
In any future philosophy of mind these must be brought into question. Here I shall merely note them, with brief comments.

The first unquestioned point of reference concerns the related concepts "physical" and "material". It has been assumed all along that these were clear and unequivocal, and that the task of materialism is to find a way in which mentality can be slotted into the material world (or else rejected as a square peg in a round hole), where we are assumed to have a clear idea of what constitutes the latter. It now seems to me, however, that physicality and materiality are many-faceted concepts, covering a diversity of phenomena which cannot, save in a question-begging way, be reduced to a single type. (The same holds good, mutatis mutandis, for mentality.) As a result of this, the issue about whether mental states are (reducible to) material things seems much more complex than I at first believed.

The second, and more fundamental, point is the notion that the basic function of mental ascriptions is correctly characterised as stating facts. I questioned this somewhat in Chapter 6, but I have for the most part let it be assumed to underlie both my views and those of my opponents. (No doubt this is because I felt I lacked the tools to make clear an alternative notion.) I now feel that this second notion should be confronted by a dilemma: either it carries with it a
particular account of what kind of fact (e.g. physical, phenomenal, behavioural) is asserted to exist by a mental ascription, or it does not. In the first case it is false, since mental facts are not reducible either to physical facts or to behavioural or phenomenal facts; in the second case, I admit that the claim that mental ascriptions are used to assert facts is true, but I must point out that this is not an explanation of the use of mental ascriptions, since whenever we regard a mental ascription as true, we will think it correct to say "It is a fact that ...": the proposed explanation is question-begging. There is the further point that mental ascriptions in particular very often have the force not of "This is how things are" but of "This is how we should regard them".

For these reasons I am led to rely more and more on the criterial account of mental language as providing a more informative explanation of its use than its historical predecessors, as well as a more accurate picture of how we actually operate with mental language. In my past thinking I restricted the criterial view so that it applied only to mental language, this for the sake of avoiding objections from a wider base. It seems true to me, however, that the view has a much wider application and that, once understood correctly, it provides the basis for a sweeping revolution in theory of
meaning. My conviction is that a specific account of the criteria for the use of an expression, as opposed to the observation that "P" asserts the existence of the fact that P, tells us what we want to know when we fail to understand the meaning (use) of the expression.

These three basic assumptions, then, are the ones which I should bring into question in any future work in this area. If my attempt is successful, it will not bode well for materialism, but neither will it for phenomenism. What is required in the philosophy of mind is not the deciding of the issues between these two hoary positions, but their transcending. I shall consider my present work successful if it indicates, to however small a degree, how this might be possible.
Footnotes to Chapter 7

1. I feel it worthwhile to mention that from this conception of the friendly coexistence of different systems of explanation one can generate a prima facie argument against Lewis' form of the contingent identity theory. According to this, mental terms are introduced to refer in advance to whatever is eventually discovered to fulfill a certain causal role (in a complicated nexus of interlocking causal relationships). The implication is clearly that the "whatever" will be uniquely realised. However, if there can be more than one (compatible and successful) form of explanation for behaviour, then there is no single answer to the question of what this "whatever" is, and therefore no reason to suppose that neural states or facts will be allotted preeminence in filling this hole. If, furthermore, the hole can be filled by states or facts described in ordinary mental terms, the identity theory becomes rapidly trivial.

2. The view is slightly overstated here, for of course the denial of the existence of mental states is, on that view, contingent upon the theoretical triumph of neurophysiology, a qualification which I have omitted for ease of expression.

3. This point was made by, and much of my discussion of reasons and causes stems from conversation with, John Westwood.

4. Throughout this paragraph "entities" may be taken to refer collectively to entities, properties, states or processes, even facts. This gives a sufficiently general reading of the "basic tenet" of reductionism.

5. Here I am drawing upon the general approach to such matters suggested by Putnam [23].

6. This is to contrast mental states with "dints", "sakes" and other cases where a noun is used in a non-referring way. See Dennett [47].

7. Chapter 5, section (b).

8. As argued, for instance, by Sikora [48]. His view rejects incorrigibility but preserves foundationalism. My own view rejects foundationalism (and incorrigibility in anything but an epistemologically useless form) but preserves certainty,
which is given a Wittgensteinian treatment.

9. My thoughts on this are obviously influenced by Ryle [49].

10. I believe the suggestions of Hellman and Thompson [43] [44] to amount to a view of this kind.
References

Abbreviated reference to a number of collections is made in the following way:


[22] R.I. Sikora: "Rorty's Mark of the Mental and His Disappearance Theory"; Canadian Journal of


[27] J.A. Black: "The Criterial View of the Mental"; see Appendix.


[34] E.S. Shirley: "Rorty's 'Disappearance' Version of the Identity Theory"; Philosophical Studies LXV (1974); pp. 73-75.

(1979).


Bibliography

Aside from the works to which explicit reference is made, I have found the following background material useful:

Eliminative Materialism


Contingent Identity Theory


[69] N. Malcolm: "A Rejoinder to Mr. Sosa"; Dialogue III (1965); reprinted in [Borst].


[77] E. Sosa: "Professor Malcolm on Scientific Materialism and the Identity Theory"; Dialogue III (1965); reprinted in [Borst].


Functionalism


[84] H. Putnam: "Minds and Machines"; in [Hook].

[85] H. Putnam: "Robots: Machines or Artificially Created Life?"; The Journal of Philosophy LXI (1964); reprinted in [O'Conn].

[86] H. Putnam: "The Mental Life of Some Machines"; in H-N. Castaneda (ed.): Intentionality, Minds and Perception; Detroit, Wayne State University
Press, 1966; reprinted in [O'Connor].


Criteriology


[92] A. Donagan: "Wittgenstein on Sensation"; in [Pitcher].


General

[101] D.M. Armstrong: "The Nature of Mind"; Inaugural Lecture of the Challis Professor of Philosophy at the University of Sydney, 1965; reprinted in [Borst].


Appendix: The Criterial View of the Mental

(a) What are Criteria?: Five Views

This question has been asked and answered in many ways. The account which I shall present is the one which seems to me both the most plausible and the most consistent with Wittgenstein's remarks about criteria. Unfortunately it has the disadvantage of being virtually unheard of in the literature. My first task then will be to dispose of the popular interpretations of Wittgenstein.¹

I shall require a number of yardsticks against which to measure these conflicting accounts. Among them will be some key remarks by Wittgenstein, which will figure in the test of the plausibility of these accounts relative to the interpretation of his later work, but I shall need to marshall some general intuitions about meaning and evidence if I am to pass judgment on their absolute plausibility. While I thus operate under a dual criterion of adequacy, I shall endeavour to give priority to the question of the absolute plausibility of these accounts and of my alternative. In this context, at least, the truth is more important than the name of him who uttered it.

The crudest of the above interpretations, which I call
the E-view, consists in the claim that the fulfillment of a criterion (either on its own or in conjunction with some statable set of background conditions) entails the presence of that for which it is a criterion. I am not sure that this view has ever been seriously entertained, but it has been attacked, for example by Rorty [99] and Canfield [89]. The E-view is seriously inadequate as an interpretation of Wittgenstein. Apart from its being contrary to the whole tenor of his thought about meaning, which seeks to undermine the importance of accounts of meaning in terms of logically necessary and sufficient conditions, the E-view is inconsistent with some of the most basic examples of criterial relationships provided by Wittgenstein. Thus the fact that there is water falling from the sky does not entail that it is raining, although in most circumstances it is a criterion of rain; the fact that writhing, screaming and moaning are criteria for the ascription of pain to others does not mean that if anyone writhes, screams or moans s/he must be in pain. A general appeal can be made in cases like this to the theoretical possibility of our discovering some further fact which will lead us to reverse any particular judgment based upon criteria.

The interpretation which might be viewed as the opposite of the E-view I call the C-view, according to which
there is no special relationship between a criterion and that for which it is a criterion, over and above the contingent relationship between a piece of evidence and that for which it is a piece of evidence - in each case the relationship between indicator and indicated is established by some form of explanatory induction. This view, common to Putnam [29], Rorty [99] and other scientific realists, amounts to a denial that there are any criteria in the sense suggested by Wittgenstein as antithetical to the notion of symptom. Water falling from the sky and the falling of the barometer are both evidence for rain, but only the former is a criterion. Hence the C-view is not that of Wittgenstein.

Both the E-view and the C-view are attempts to reduce the criterial approach to meaning to more standard views. One cannot show simply that these standard views are false, but I hope to add plausibility to this contention by the development of the criterial view as an alternative.

The most popular interpretation in the early post-Wittgensteinian literature, which I shall call the N-view, is an attempt to cast the criterial view as an alternative to these standard theories, while preserving the best elements of each. The view, held by Kenny [94], Lycan [95] and Shoemaker [100], consists in the claim that to say that X is a criterion for Y is to say that it is necessarily true that X is evidence
for Y, with the result that if X is a criterion for Y, it is logically impossible for X to come to fail to be evidence for Y. The N-view has perhaps gained currency by being confused with the claim (with which, by appeal to the most common of modal fallacies, it is interderivable) that it is a necessary truth that if X is a criterion for Y, X is evidence for Y. To the (quite large!) extent that this claim is undeniable, it also appears compatible with any reasonable account of criteria, including all those here to be rejected, and so can be ignored.

The N-view has been refuted by Hilary Putnam [29] on the perfectly acceptable supposition that, if anything is a criterion for anything, pain-behaviour is a criterion for pain. Putnam asks us to consider a society of congenital stoics who never manifest pain in their overt behaviour, but to whom it is in principle possible for us to ascribe pains on the basis of neurophysiological evidence. While I think the description of such a community is a taller order than does Putnam, I feel that such a situation as he describes is perfectly conceivable. Here would be a case, then, where pain-behaviour has ceased to be evidence for pain; this constitutes a counter-example to the N-view.

A defender of the N-view might object that s/he is committed not to the thesis that (in the case of pain)
behaviour always will be evidence for pain, but to the thesis that it always would be evidence for pain, were it to occur. Its non-occurrence in the described state-of-affairs does not invalidate its potential evidential status. Although Putnam does not seem to anticipate this objection, it can be handled without much difficulty by extending his description of the society of stoics. Thus we have only to suppose that the stoics do exhibit pain-behaviour, though not in response to pains, to evade the force of this counter-objection.

There is a more general defence, publicised by Norman Malcolm [96] [98], to criticisms of the N-view based upon counter-examples like Putnam's. These examples rely on the supposition that neurophysiological evidence might become better evidence for the occurrence of certain mental states than overt behaviour. In Malcolm's view, this amounts to the acquisition by the neural evidence of criterial status and thus to a change in whatever mental concept we are concerned with, a change in the meaning of the relevant mental term. If this is so, then the fact that one can use the same word to ascribe a "new" state on the basis of a new sort of evidence is not a counter-example to the claim that it is necessarily true that the old sort of evidence is the basis for the ascription of the "old" state.

The best argument against this defence is also
presented by Putnam, this time in "Dreaming and Depth-Grammar" [23]. Putnam considers the case of a term, "multiple sclerosis", at present used to ascribe a disease to a person on the basis of a set of readily observable symptoms (in the medical, not Wittgensteinian, sense). Let us assume, as he does, that there is a justified presumption, borne out by the facts, that such a cluster of symptoms does pick out a single condition with a single, unknown, viral cause. Let us suppose too that in fact doctors have never been wrong in ascribing this disease, that there were no cases where the symptoms were present, but the unknown cause not, and where such a diagnosis was made. Let us make the corresponding assumption for the correctness of refusals to diagnose multiple sclerosis. In such circumstances I find no difficulty in asserting that the connection between symptom-cluster and disease is criterial.

Suppose now that the virus resposible is discovered, and that future diagnoses are based entirely on the presence and activity of this virus in the patient's body. Once this discovery has been made, certain patients displaying the cluster of symptoms will be declared not to have multiple sclerosis, on the grounds of the absence of the virus: in other words, the criteria have changed through normal medical practice. It seems to me wrong to hold, as the N-view entails, that a new concept of multiple sclerosis has come
into existence, or that the term "multiple sclerosis" has changed its meaning. On the assumption that the relevant connection is criterial, then, the N-view can be rejected.

One more account of criteria remains to be considered before I move on to the exposition of my view. This I call the V-view, which consists in the claim that criteria are ways of verifying (sc. settling with non-logical certainty) statements about the presence of some item. It is not clear that this view is incompatible with either the C-view or the N-view (the qualification implied in "non-logical certainty" rules out the E-view) and indeed Malcolm holds both this and the N-view [96]. The V-view stresses the epistemic as opposed to the quasi-definitional role of criteria but it is this very difference of focus which allows for the compatibility. It is to be found in the accounts of criteria due to Canfield [89] and Albritton [88]. As long as it is distinguished from the E-view, the V-view is acceptable - it says something true about criteria. My only criticism of the view is that what it says can also be said about any form of strong evidence, and so does not provide a description distinctive of criteria, nor therefore an analytic account of what criteria are, unless much more be said. ²

My own view, which I shall call the G-view, takes as its slogan the claim that the criteria for the use of an
expression are the grounds upon which one would justify the use of that expression. Hence a criterion is a reason. Now this by itself is again not enough to distinguish my view from the C-view, for if I have evidence for a person's being in a certain mental state, then I have a reason for asserting that s/he is in that state. We need to focus more sharply on the sort of reason that is relevant and on some of the circumstances attending the ascription of the relevant state before we can make the distinction between criteria and evidence.  

One is tempted to begin by stressing the logical, as opposed to empirical, nature of the connection between criterion and that for which it is a criterion. The connection is supposed to be one of meaning, not one established by discovery. It is conceptual, not factual. The trouble with these formulations is that they invite us to adopt either the E-view or its watery cousin the N-view, and I have argued that these are incorrect. What can I propose in their stead?

Unfortunately there is no short answer to this question. The account which I have to offer does not take the form of a continuation of the assertion "The connection between criterion and that for which it is a criterion is ...". (To this extent the Question "What is the connection?"
resembles the question "What is time?". What is needed here is not a definition, but a description of some of the ways in which we operate with criteria, for we need to be reminded of a phenomenon which is already familiar to us in ordinary life, as much as it resists representation within the categories of standard philosophical analysis.
Let us concentrate first on the role of criteria in explanations of meaning. Most of the contexts in which Wittgenstein himself mentions criteria (in the Blue Book, Brown Book and Philosophical Investigations) concern the disambiguation of sentences (or ascriptions). If a key term in some ambiguous sentence is used in, say, two different senses, one way to clear up any confusion which might result is to specify the two senses by saying what in each case would support the judgment made by the utterance of the sentence. It is important to see that this is not verificationism. I do not say that the meaning of the sentence is to be laid out definitionally in terms of its method of verification, but that you can draw my attention to a particular sense (with which I am more or less familiar) by mentioning the method of verification which is appropriate for statements made in that sense, but not the other. The method of verification gives us referential, as opposed to definitional, access to the meaning of the sentence. This is a fairly trivial observation, and so far does not indicate any important distinction between criteria and evidence, since you may achieve the same feat of disambiguation by mentioning something that would count as
evidence for the judgment taken in one sense but not the other. (What I mean is that what you do may equally well be described as "mentioning evidence").

The circumstance imagined above is that in which a method of verification is appealed to in order to tell me which of two familiar senses of an expression is intended. In this way you might try to show me that when you said "He fell off the bank" you meant a fluvial and not a financial bank. There is a different sort of circumstance to be considered too, and that is one in which the meaning of an unfamiliar expression is to be explained to me - here I do not have the background knowledge necessary for the success of the "referential" approach. In some cases, where I have sufficient grasp of other parts of language, you may attempt a definition in terms of the necessary and sufficient conditions for the application of the expression; in some cases an ostensive definition may even be possible. In the vast majority of cases, however, these resources are not available - there simply is no definition of "chair" or "belief in the omnipotence of God". (If at this point you are inclined to stop and try to formulate one, reflect at least that anything you are likely to come up with has not in the past been used to explain the meaning of these expressions to you.) What we fall back on, within this majority of cases, are features of
the potential referents of the expression which indicate (without entailing the claim) that the referents are counted as falling under the term.

Several questions at once arise at this point. If one is inclined to regard the above indicators as connected evidentially with the referent of the expression, it is useful to introduce a distinction between strong and weak evidence. (Nothing I shall say is incompatible with the fact that the difference behind this distinction is one of degree.) There are two reasons for this, the second of which is more important to the conception of criteria. The first reason is that, generally speaking, it will be more efficient, in explaining the meaning of an expression by this method, to refer to indicators which constitute strong evidence for the occurrence or presence of the referent of the expression. This is because the weaker the evidence is, the greater is the number of alternative hypotheses which could account for it and so the greater is the chance that the hearer will mistakenly pick one of these as the fact asserted by the use of the expression.

The second reason emerges when we ask under what conditions we assert that the hearer has successfully understood our explanation. Here it seems correct to say that we shall be more inclined to deny that s/he has understood if
in the future s/he discounts strong evidence for the occurrence of the referent of the expression than if s/he discounts weak evidence. For, since the number of alternative explanations of the weak evidence is greater, so is the number of ways in which we can ascribe to her or to him alternative beliefs about the explanation of the relevant evidence. Conversely we shall attribute to her or to him an understanding of the expression if, in general, s/he is inclined to use it in accordance with the prevailing view of what is strong evidence for the occurrence or presence of its referent. 4

Now it is this distinction between strong and weak evidence that I wish to replace with that between criteria and symptoms, the latter term covering all that is properly regarded as evidence. How the new distinction differs from the old is not easily shown, and the attempt will be postponed for the moment. This is of course the pivot around which the criterial account of the mental turns.

The point of my rehearsing the part of the criterion/symptom distinction using the evidential distinction as understudy is to confer plausibility upon the idea that facts which are normally regarded as contingently connected with the occurrence or presence of the referent of an expression play an essential role in the account of the meaning of the expression, and of what it is to understand it
and use it correctly. There is no other way of explaining the meaning of most ordinary terms.

As a result of the arguments of Quine and others, it has become commonplace to assert the impossibility of distinguishing between those features of the referent of an expression which it has in virtue of the meaning of the expression and those which it is merely commonly believed to have. I suggest that this conclusion can be replaced by the claim that the relevant distinction covers a difference of degree and that this can be explained in terms of the criteria for understanding the use of the expression. Certain of the beliefs are more closely connected with the meaning of the expression (which means that we shall be more inclined to deny that anyone who, without proper reason, rejects them has understood the meaning) but none is logically tied to it, if the latter notion is to be explicated in terms of entailment.

Consider the following passages from Philosophical Investigations:

353. Asking whether and how a proposition can be verified is only a particular way of asking "How d'you mean?". The answer is a contribution to the grammar of the proposition.

354. The fluctuation in grammar between criteria and symptoms makes it look as if there were nothing at all but symptoms. We say, for example,: Experience teaches
us that there is rain when the barometer falls, but it also teaches us that there is rain when we have certain sensations of wet and cold, or such-and-such visual impressions." In defence of this one says that these sense-impressions can deceive us. But here one fails to reflect that the fact that the appearance is one of rain is founded on a definition.

355. The point here is not that our sense-impressions can lie, but that we understand their language. (And this language like any other is founded on convention.)

One of the important points being made here is that anyone who, in normal circumstances, did not understand that certain sense-impressions provided us with a reason for asserting that it is raining would not be said to understand the sentence "It is raining". This is not an exceptionless rule - a philosophical sceptic might fail to pass by this standard - but it is significant that a special explanation is required before exceptions are or can be allowed. Criterial rules are generally of this nature. The "logical" character of the criterial relation resides, in so far as it exists, in the fact that using the term according to its conventional criteria is a criterion, and a very important one, of understanding the term.
The role of criteria in explanations of meaning is derivative from their more basic role in the justification of assertions. (Here I mean the justification of real assertions made in ordinary life and not the epistemological project, of doubtful origins and dubious validity, known as the Cartesian Quest for Certainty.) I now turn to this context in furthering my attempts to drive a wedge between criteria and evidence.

The first point of note is the fact that criteria justify, as it were, without further ado. By this I mean not only that people do not, as a matter of fact, request further support for the criterial justification of an assertion, but that in an important sense none is available. In general there can be no explanation (save the trivial semantic one) of why the fulfillment of such-and-such criteria is grounds for such-and-such an assertion. This situation contrasts with that in the case of evidence, where the relation between particular piece of evidence and particular hypothesis is mediated by an explanatory theory. A development of the rain/barometer example of Wittgenstein's #354 (quoted above) provides an excellent illustration of just this point. The
connection between the falling barometer and rain can be explained by meteorological theory. I have some idea of what could count as an explanation of the connection between "certain sensations of wet and cold, or such-and-such visual impressions" and rain, but it is hard to imagine a practical case where this explanation would be necessary. What if one asked for an explanation of the connection between water falling from the sky and rain? (This last connection I call criterial.) One is tempted to reply to the imagined request for such an explanation: "But that is just what rain is", or "But that's what the word means"; one would however be wrong if one took these last statements as imputing to the relevant criterial connection the character of entailment. Water falling from an aeroplane is not rain. One would also be wrong if one felt the absence of the relevant explanation as a lack: none is available, but none is required. Our use of words is such that this is the case. I shall argue later that this feature of criteria-governed language has an interesting ontological consequence.

Now it might be objected that the line of thought I am here following relies too heavily on a restricted concept of evidence. Not all evidence is of the theory-mediated sort. Even in the above example it could be that the connection between the falling barometer and rain has been established by
simple induction, as a constant concomitance of events. There may be (and presumably are) other sorts of evidence still. Am I not being artificial in arguing for a distinction between criteria and evidence on the grounds that criteria are not like one sort of evidential connection? My attitude to this objection is to accept it, and thus clarify my position. Yes, there are different kinds of evidence - there may for all I know be a sense of "evidence" in which criteria as well as the kinds I wish to distinguish them from are all correctly called "evidence". My aim in this section, however, is to mark the criterial view as different from and superior to theoreticalism in the philosophy of mind, and the forms of this view which I consider in the main body of the thesis all represent the evidential connection between behaviour and mental state as being of the theory-mediated kind. The wedge I am driving, then, lies in the crack between criteria and theory-mediated evidence, as will be apparent from what follows. Once I have forced open this gap I can hope to show on which side of it the mental case falls.
I should like now to propose an example of a criteria-governed term which instantiates the features I have been discussing as distinctive of such terms. The example will enable us to see how such a term can operate in the absence of even an implicit theoretical explanation of the connection between the truth of the statements employing the term and those used to justify or support them.

The term I have chosen is "danger". I am interested mainly in attributions of states of being in danger, since these provide ready analogues of ascriptions of mental states, so I shall consider statements of the form "X is in danger". How are such judgments supported, and what is the nature of the connection between supporter and supported?

The greater part of supporting statements in this case are circumstantial: the circumstances in which X finds itself are such as to pose a threat to X's existence or integrity. We will usually mention such circumstances in support of our danger-attributions: Perseus is in danger because he is fighting Medusa; Holmes because the cliff-edge is crumbling away; the house on the corner because its demolition is planned. In the case of sentient beings who are aware of the
danger, we can also have behavioural grounds, in their attempts to escape whatever situation they are in, manifestations of fear etc., for a danger-attribution.

Now it seems to me obvious (though it may not to the reader) that no sensible definition of "danger" could be formulated which would make the connection between danger and either of these two forms of ground, circumstantial and behavioural, one of entailment. (This familiar point is again one for which I shall not argue.) This much the two forms have in common; beyond it they differ radically.

The relationship between behaviour and danger seems to be straightforwardly evidential. We explain his trembling by his fear, and his fear by his being in danger. Our evidence consists in the effects of danger, the posited explanation, and the mode of explanation is straightforwardly causal. Such is not the case, however, with the circumstantial grounds, for these do not provide us with evidence of the existence of an independent state-of-danger in which the subject stands. I should prefer to say that they constitute that state. Perseus' danger consists in nothing other than the fact that he is fighting Medusa, and that she is a terrible enemy. A grammatical indication that the grounds for a state-attribution constitute the state, as opposed to acting as evidence for its existence, is the substitutability of the
intentional conjunction "in so far as" for the apparently "epistemic" "because"\(^9\) in typical statements which reveal the justification of the state-attribution. "Perseus is in danger because he is fighting Medusa," but also "Perseus is in danger in so far as he is fighting Medusa". "Perseus is (must be) in danger because his face bears a look of terror", but not "in so far as his face bears a look of terror" (though circumstances could be imagined in which this too were possible).

For reasons such as these I wish to say that the relationship between circumstances and danger is criterial, and that between behaviour and danger evidential (or symptomatic, to use Wittgenstein's terminology). The picture I have drawn of the first relationship enables me to distinguish quite easily between criteria and theory-mediated evidence, for a theoreticalist account of this relationship has little plausibility. It is permissible to say (informally) that the presence of Medusa causes Perseus to be in danger, and that this state of danger causes Perseus' fear, but it would be wrong to conclude from this that danger is a theoretical entity hypothesised as a causal intermediary between circumstances and behaviour. The first cause mentioned here is constitutive, not efficient, and as such might itself be said to be the (real) cause of the fear. Furthermore, when we
Furthermore, when we explain the fear by supposing that the subject is in danger, we are not positing a theoretical entity, "danger", which, according to a concurrently hypothesised law, causes fear - we are merely supposing that the subject is (aware of being) in some situation (within a more or less familiar range) which is potentially detrimental to his, her or its wellbeing. The state of danger goes proxy, as it were, for a disjunctive cluster of circumstances.

This point can be made clearer by the introduction of two distinctions among hypotheses. By **theoretical hypothesis** I mean the positing of theoretical entities in causal explanation. By **nontheoretical hypothesis** I mean the forming of an empirical generalisation about entities whose existence is no longer a merely theoretical matter. (E.g. "Smoking causes lung-cancer.") By **subhypothesis** I mean the supposition that some particular case instantiates a particular general hypothesis. The artificial theoreticalist account of the meaning of "danger" takes danger to be a theoretical entity, and particular danger-attributions to be theoretical subhypotheses; on my account the reference to danger in the causal explanation of fear serves to introduce an (indefinite) disjunction of nontheoretical causal subhypotheses, which instantiate various nontheoretical hypotheses causally connecting various particular circumstances with fear (and
ultimately behaviour). I shall have more to say later about the relevance of these distinctions to the philosophy of mind.

My account of danger should not be taken as having the consequence that danger does not exist. Danger exists if and only if something is in danger, and there are conditions under which this last statement is true. All I have argued is that the relation between these conditions and danger is not definitional; that the conditions provide reasons which support the danger-attribution; and that the nature of this support is other than that of evidence for hypothesis. I note too that we are not inquiring here into the reasons which underscore the attribution's moral, aesthetic, economic etc. value - what is at issue is the linguistic appropriateness of the danger-attribution. Reasons of this sort I call criteria.
(e) Criteria and Mentality

The point of the "danger"-example is to motivate the making of a distinction between criterial connection and the relation between evidence and theoretical hypothesis. I believe it succeeds in this task. However, there are important disanalogies between the cases of danger and of the mental, which must be explored. The most obvious one is that whereas the constitutive grounds of danger are completely circumstantial, those for (many, at least) mental states are both circumstantial and behavioural. I think it probable that detailed analysis would show that the proportions of each change from one mental state to another - grief is more circumstantial than depression, understanding more behavioural than perceiving - but I make no pretence of such analysis here.

I am sure that some reluctance will still be felt about calling behaviour a constitutive ground for any mental state - this reluctance is fostered by concentrating on the so-called "occurrent", or at least so-called "non-dispositional" states, such as sensation and perception, for it does seem quite plausible to regard, for example, speaking French as constitutive of understanding French in certain cases. With the occurrent states, however, the difficulty remains. A
point which seems strongly to favour reluctance is that it makes perfect sense to ask "How do you know he is in pain?" and to receive the reply "Because he winces and holds his arm". The clearly epistemic character of this exchange makes it look just as clearly as if pain and behaviour are two things, not one.

This, however, is not denied, though my stress on the notion of "constitution" might make it appear otherwise. Although I think that many of the criterial grounds for the ascription of mental states can justly be called constitutive — dispositions to behave are a good example in the case of "dispositional" states — the notion of criterion is wider than that of constitutive ground, and this is not brought out by the "danger"-example. There I claimed that the circumstantial grounds were constitutive of danger in order to avoid the idea that danger is a theoretical construct; it seems clear that the claim achieves this goal. What I need to provide in order to achieve the same goal with respect to mental states is a nontheoreticalist picture of the relation between criterion and mental state in the case where the criterion cannot plausibly be regarded as constitutive of (part of) the mental state.

Wittgenstein remarks that "'Grief' describes a pattern which recurs, with different variations, in the weave of our
I propose to build my account around an extension of this conception to all mental states. The characteristic circumstances for the occurrence of a mental state, and the behaviour which typically manifests it, (when none of these is part of the state), are merely parts of the pattern, not the whole. Nevertheless, they are the parts which function in the identification of the pattern as familiar (as similar to previously-encountered cases) and thus in the ascription to the relevant subject of the relevant mental state. This is what their criterial role amounts to in this case.

Some qualifications must be made immediately. First comes the by now familiar point that behaviour and circumstances are not definitionally tied to particular mental states. So it does not follow from the occurrence of the circumstantial and behavioural parts of the pain-pattern that the subject is in pain. Secondly, it is misleading, though convenient in what follows, to speak of the pattern for any particular state, as this implies uniqueness. Although it is by no means clear how in this context one is to count patterns, it seems more natural to say in general that there may be a number of patterns, connected by bonds of similarity and family resemblance, associated with each mental state. This is peculiarly obvious in the case of "propositional
attitudes", for while all belief-patterns may be similar, the belief-that-p-pattern must differ significantly from the belief-that-q-pattern. Even this formulation is unsatisfactory in so far as there may be many circumstantial/behavioural patterns associated with as specific a belief-that-p as one cares to consider. These features of my picture make the identification of mental states a very complicated business, but this is not, I think, a disadvantage. It seems to me that the recognitional capacities picked up in learning mental language are every bit as complicated as my account suggests.

The third qualification concerns the nature of the patterns I am discussing. There is a temptation to argue as follows: "Even if we accept the pattern-account, we can still ask, 'What is the mental state?'. There seem to be only two possible answers: (i) it is the whole pattern, and (ii) it is merely part of the pattern, specifically that part which is neither circumstantial nor behavioural." The first answer leads us in the direction of behaviourism and is to be rejected - it is a presupposition of this discussion that wincing, for example, is not part of pain. The second alternative seems to push us towards a theoreticalist view of the mental: we recognise a pattern, as it were, by its beginning and end (circumstances and behaviour), and
hypothesise a mid-section, to which we give the name of the relevant mental state. If this is the way the notion of a pattern is to be elaborated, the criterial view collapses into the theoreticalist one.

In order to distinguish between the two views, and to show the superiority of the criterial one, I shall make two points. First, it is a presupposition of the theoreticalist account that the relations between elements in such patterns are predominantly causal: the mental state is posited as a causal intermediary between circumstances and behaviour. On the criterial view, this is not necessary—situations can be recognised as instantiations of a particular pattern purely on grounds of similarity. Even if the relevant similarities include familiar concomitances, these need not (though they may) be regarded as causal in character. For example, to recognise that someone is in pain after he has touched a hot-plate, I do not have to make the supposition that the heat of the plate caused (by way of some intermediary) his wincing—I may simply notice that this situation (touching, sudden withdrawal and wincing later in time) is one which is familiar to me and in which I have been taught to ascribe pain, barring the discovery of further facts which might speak against this ascription.

Secondly, it does not make sense, on the criterial view,
to identify the mental state with any part (including the whole) of the pattern. All we know is that when such a pattern occurs, we are entitled to ascribe a mental state to a subject. Compare this with the case of another "immaterial object": when certain patterns of trade and other human interactions occur, we are entitled to say "The economy of Canada is booming", but the question "Which part of those patterns is the economy of Canada?" has no sense (or none as yet). Depending upon what mental ascription is at issue, different parts of the patterns will be given more weight in supporting the ascription. For instance, the recent loss of a loved one is a circumstance highly relevant to my ascription to you of grief (the closest thing to a necessary condition, in fact), but while the circumstance is part of the relevant pattern, it does not constitute the grief (even in part).

The third qualification is, then, that the patterns I am discussing are not causal patterns involving circumstances, mental state and behaviour, but are the patterns of events and states other than the state-to-be-ascribed, which we (could) appeal to in the justification of our mental-state-ascriptions. When I say that in any case only part of the pattern is available to us, I mean only to leave open the possibility that facts of a different sort may come to be relevant to the justification of such ascriptions. The prime
example in the present context are the neurophysiological facts about the subject.

To see how these might become criteria for the ascription of mental states we must develop the present account somewhat. In the course of this development I shall try to show how the criterial account substitutes a certain kind of certainty in mental ascriptions for the alleged hypotheticality of mental states. I shall begin with the role of what I have (rather clumsily) called nontheoretical subhypotheses about another's mental states.

Consider a case in which I have a relatively small amount of knowledge about a subject's circumstances and behaviour and wish to describe her/his mental state. On the present account, what is available to me is a small number of elements of some pattern. Now it is in the nature of the case, when we are speaking of the mental realm, that elements can occur in many different patterns: the death of a family-member figures in grief, revenge, the anticipation of insurance benefits and so on; wincing may signify pain, sudden realisation of danger or error, disapproval of a poor joke and so on. Consequently, one cannot tell, from observing a small portion of a pattern, what pattern it is, nor therefore what is the mental state of the subject. It is possible, however, to make a supposition about what it is - on my view a
nontheoretical subhypothesis (but I must not beg the question against theoreticality at this stage). What one does here is to reconstruct the greater part of a pattern from the lesser, and this reconstruction may turn out to be accurate or inaccurate, and the resulting ascription true or false.

Now since the patterns associated with any particular mental ascription are diverse, and each contains a number of elements, there are many different combinations of small numbers of elements which can prompt us to make any particular subhypothesis. Since in the case of each possible combination we say we explain the behavioural part of the combination by the subhypothesis we make, this gives the illusion that the mental ascription is a theoretical subhypothesis called in to explain causally these diverse effects. On my view, however, such is not the case. The mode of explanation displayed here is not theoretico-causal in any simple sense. What I do when I explain a piece of behaviour by referring to the subject's mental state is to place the behaviour within a recognisable pattern, to represent it as part of this pattern rather than that. This is explanation in the sense that, once you have it, you know more about the case at hand than you did before. Since some of the relations between pattern-elements are causal, you will know more causal facts with the explanation (assuming it is true) than without it; furthermore, if you
have a true explanation you will be allowed to express it in the form "his wincing was caused by his pain" (for example). Thus mental explanations license causal judgments of various types, but it does not follow from this that mental states are theoretical entities.

One way to see this is to notice that the description of mental explanation relies on the concept of familiarity. The explanation of a particular case relies on relating this case to cases previously encountered: in general one will understand such an explanation (delivered verbally) only if one is already familiar with patterns that are relevantly similar. Acquiring such an understanding is a process of familiarisation. This feature is not characteristic of theorectico-causal explanation. (One must familiarise oneself with the mode of explanation, the operation of the theory, in order to understand the theory, but not necessarily with the range of cases to be explained. 10) These considerations about familiarity do not entail, however, that "new" mental states are impossible (or incomprehensible) - new beliefs, for instance, arise every day. New states acquire their names, as it were, in virtue of the similarities that hold between their patterns and old patterns. We extend the range of patterns associated with the different types of mental state, and we extend the list of types of mental state (most commonly those
whose ascriptions involve intentional constructions) by reference to the presently-existing ranges and lists, and largely in accordance with the normative constraints of rationality.

If the above account of mental explanation is correct, some interesting results ensue. One is what one might call the "pervasive ambiguity" of mental utterances, were it not for disanalogies between this characteristic and other forms of ambiguity. What I mean is this: I can explain many different pieces of behaviour (and possibly other sorts of pattern-element) by use of the same mental ascription. Here the ascription roughly has the force of "The rest of the (relevant) pattern obtains", but of course "the rest of the pattern" covers a different portion of the pattern as the part-to-be-explained is different. So in different contexts (where explanation is the issue) mental ascriptions will have different "meanings" (i.e. will convey different pieces of information) depending upon how much the hearer already knows (e.g. what s/he is seeking to explain). In each case, however, the general aim of the explanation is to fill in the whole picture.

In the above discussion of the role of mental explanations I have concentrated on cases where only a small part of the typical pattern is available to observation, and I
have tried to distinguish this mode of explanation from the theoretico-causal mode, while at the same time trying to display those features which ground the illusion that the theoreticalist account is correct. Now I wish to turn to cases where a much larger portion of the pattern is available to observation. The more one increases the size of this portion, the more one restricts the range of possible interpretations of the available pattern (sc. number of standard patterns one can plausibly represent it as part of). It is my contention that while no pattern-description, no matter how full, entails that the subject is in a certain mental state, the corresponding mental ascription can be made with certainty on the basis of the sum of those parts of the pattern which are readily observable to an outsider, i.e. circumstances and behaviour. To explicate this contention I need to develop a brief account of certainty as it occurs in this context.

It should be clear that the certainty I propose is of the "non-logical" kind also proposed by supporters of the V-view, discussed earlier. Both the criterial and theoreticalist views reject the (behaviourist) claim that the results of the observation of circumstances and behaviour entail the mental facts about a person. Both agree, then, that ordinary subhypotheses about another's mental states are
to this extent provisional: while the accumulation of circumstantial and behavioural facts gives us more and more reason for a particular mental ascription, they are never sufficient to put the ascription beyond logical doubt. This is because any ascription is potentially revisable (as far as we know) in the light of future discovery (e.g. in neurophysiology). However, the fact that there is a logical possibility of error in making a mental ascription does not entail that I am in error when I make it, or that I do not know the truth of the matter. Supposing that the ascription is true, I will know it (in most ordinary cases) if I believe it on the right sorts of grounds. Both views agree on what these grounds are - circumstances and behaviour, at the present time, - but disagree on the nature of the epistemic connection between the grounds and the ascription.

Here, as earlier, the theoreticalist account divides into two. On the one hand we have the identity-version, as expressed by Lewis [14] [15], on the other the eliminative version. On the latter account, mental ascriptions are subject to a deeper form of doubt than attends the already-admitted logical provisionality of subhypotheses, for there is a theoretical doubt about the very existence of mental states, which if fulfilled would render all mental ascriptions false. If we hold the identity-version, we are saved from
this possibility by the provision that mental terms have been introduced (as it were, in advance) to refer to whatever causes behaviour (in certain ways); on this theory, while there can be doubt about whether a particular subhypothesis is correct, on a particular occasion, there can be no doubt about the set of theoretical hypotheses which mediate the evidential connection between behaviour and subhypothesis.

It is clear that, of the two theoreticalist views, the identity-version more closely approaches the criterial account of the matter, in so far as it places "mind-theory" in a sacrosanct position. It is here relevant to dispute only Rorty's eliminative version.

The central contrast between this and the criterial view is of course the intuition, on the one hand that the existence of mental states is beyond dispute, and on the other that it is highly questionable. I do not know how to argue for the former intuition (mine) from any neutral ground, but I can say why I favour the criterial view. The eliminative version, in the form espoused by Rorty, relies on characterising mental states in a certain way (as involving incorrigibility) and then underscoring the problems associated with such incorrigible states in order to support doubt about their existence. This view relies on a theory of meaning for mental terms (however much the notion of "meaning" is eschewed
by its proponents), and this is where the criterial view diverges. As I have argued in Chapter 3 of the thesis, the ordinary concept of the mental does not have such "Cartesian" implications: in particular there is nothing to suggest that the existence of mental states is incompatible with there being a complete causal explanation of behaviour (described in physical terms) which refers only to physical and neurophysiological states and events.

In the absence of such an incompatibility, one cannot see mental language and neurophysiology as competing theories of behaviour, and so this way of trying to show that there is something wrong with mental language fails. There may be other ways which, however, have not entered the discussion. Now the point of the criterial view is to show us that we do not need to worry about this possibility - there cannot possibly be anything wrong with the use of mental language in general, though of course we may use it to speak falsely, inaccurately and so on, on particular occasions. This much is guaranteed by the pervasive role of the use of mental language in human life. However, if one has an incorrect theory of meaning for mental language, doubt about its general validity can indeed arise, and this is what has happened in the present case. The criterial account of the meaning of mental terms does not allow such doubts to gain a foothold, as I shall
attempt to explain.

According to the criterial view there is only one way to explain what mental states are, and that is to explain what is the meaning of ascriptions of mental states. Furthermore, there is only one way to specify the meaning of mental ascriptions, and that is to explain the criteria which govern their application. To do this is to give an account of the reasons which would support such an ascription. But these reasons are not pieces of evidence for the truth of the mental ascription — there is no independent way of specifying what the ascription amounts to, no definition, and no explanation of a theoretical kind of why these reasons support this ascription. This last contention may seem to be false when we consider a small part of the criteria for the application of some mental term in isolation from the rest — a small part of one of the typical patterns associated with this term, in my earlier terminology. We can ask why this is a reason for ascribing (e.g.) pain — the answer is that it is part of a particular pain-pattern, in other words that it tends to occur in conjunction with a number of other events, ordered in certain ways. However, there is no explanation of why the whole pattern is a reason for ascribing pain, other than the trivial "That is just the kind of thing 'pain' means".

At the same time we have to resist the conclusion that
these reasons entail the relevant mental ascription. Any subhypothesis may be revised in the light of future discoveries, no matter how full the criterial justification for it. But this is just a feature of the manner in which criterially-governed terms have meaning - it is possible that the criteria for their use might change, allowing for the revision of certain particular ascriptions. Now the strong claim I wish to make is that no future discovery will lead us to revise the greater part of the mental ascriptions that we make in such a way as to suggest that mental states do not exist. Mental language, and with it mental states, are too firmly entrenched in our way of life.

This "conceptually conservative" claim will not be popular among my materialist or reductionist opponents, but I hope to be able to justify it. Let us look first at the possibilities for conceptual revision which my "non-entailment" view leaves open. In the present case, the most relevant candidate for a new criterion for the mental lies in neurophysiology. It seems overwhelmingly likely that the neural facts about a subject correlate (to some as yet undetermined extent) with the mental facts about him/her. Translating this into the terms of the criterial view, we can say that there is a part of each particular pattern, namely the neural facts, which is in general unknown to us (in
ordinary life) when we make mental ascriptions. With the popularisation of neurophysiology (and, if necessary, the mass-marketing of cerebroscopic equipment) we may come to support our mental ascriptions by appeal to the neural facts. At first, these facts may be merely correlational evidence, but it is also possible that they acquire criterial status. This would be the case if we sometimes preferred the neural indicators to our standard circumstantial/behavioural criteria in cases of apparent conflict. I see no reason why such cases should not arise, and I agree with most materialists that there could be good reason for overruling the standard criteria (and, indeed, the subject's self-ascription) in a small number of cases. What is not possible is that we justifiably come to regard the majority of our ascriptions based on the standard criteria as false.

This follows from the considerations about belief and meaning which I mentioned earlier in section (b). A relatively small change in beliefs about mental states can be accomplished without changing the meanings of mental terms, but a relatively large one, such as the revision on neural grounds of most previous ascriptions of mental states, does indicate a change in the meanings of those ascriptions. Furthermore, it leaves it open that those ascriptions were true in their original senses, even if the latter have dropped
out of usage. (All this is the case, of course, provided that the new discoveries we make are not such as to make us suppose that the original criteria were never, or hardly ever, fulfilled - otherwise there is indeed no limit to the extent to which we may revise our list of true mental ascriptions.)

We can thus be certain that most of our mental ascriptions are true, given that in each particular case the present criteria are fulfilled. We do not know which of our ascriptions will be open to revision in the light of future discovery, but we know that they will be small in number. I see no reason, furthermore, why this potential revisability should lead us to say that we do not now know the truth of any particular mental ascription. My accounts of truth and of knowledge, as applied to criterially governed expressions, are hence roughly as follows: a mental ascription is true if (i) the present criteria for its use are fulfilled and (ii) no future discovery will lead to a change in the criteria for its use which would necessitate its being revised. A person knows (is/can be certain) that a particular mental ascription is true if (i) it is true and (ii) s/he believes it to be true on the basis of the present criteria for its use (or s/he has sufficient evidence of some kind for believing that the present criteria are fulfilled). This is what truth, knowledge and certainty amount to in this context.
The criterial view hence allows for certainty in the making of mental ascriptions without appealing to an entailment between criterion and mental state. The view of certainty briefly described here is taken from Wittgenstein, particularly from "On Certainty" [7], but it would be out of place to dwell on that epistemological dialectic in the context of the present essay. The most important conclusion is that the existence of mental states in general is guaranteed by the fulfillment of the criteria for the use of mental ascriptions; it is clear that this is sharply at odds with Rorty's eliminative version of the theoretical account. In the above line of argument I have stressed the sense in which the connection between ascription and criteria is one of meaning, not evidential support, but I have also qualified this claim heavily in order to avoid the charge of logical behaviourism.

For those unconvinced by my conceptual conservatism I make an addition to the foregoing account. This is a second line of argument against the hypotheticality of mental states. Even if I am wrong in thinking it impossible (really impossible, not logically impossible (?)) for the majority of mental ascriptions to be revised, I can still argue that no discovery will lead us to replace the majority of mental ascriptions with something other than different mental
ascriptions. Most of the standard examples (in the epistemology of mind) where we are misled about a person's mental state by her or his behaviour are ones in which either (i) the sample of behaviour, and knowledge of the circumstances, available to us are too small to count as criterial, or (ii) the correct explanation of the behaviour lies in the ascription of a different mental state (e.g. pretending to be in pain). Examples of the first type are easily handled by the pattern-account, as should be clear. Those of the second type, while they should provoke care and concentration on detail in the description of the use of criterially-governed expressions, cannot foster any ontological doubt about the existence of mental states. It will be readily admitted that cases which fall outside these types are relatively uncommon in ordinary life (e.g. where we suppose a "purely" physiological explanation), which is all that my general view requires. What I say here is true, of course, only on the assumption, already argued, that mental and neurophysiological explanations are not for the most part incompatible.
Footnotes to Appendix

1. This negative side of my account is reproduced with amendments from my M.A. Thesis [28].

2. Lawrence Resnick has helped me to see an important way to fill out the V-view. His contention is that the distinction between criteria and evidence is not a theoretical distinction to be explained in philosophical terms for all cases, but rather one which is to be made piecemeal in the specific contexts in which it arises. Thus the difference between a criterion of pain and evidence for pain may be totally different from that between a criterion of rain and evidence for rain. I am now convinced that this line is what was intended by Wittgenstein and have been led to adopt it myself. I feel that it is consonant with what I identify in the text as my view.

3. If, that is, the distinction can be made generally. I feel some attempt can be made in this direction. If the relativity-to-context referred to in fn. 2 above is a problem here, my account of the distinction can only be taken as applicable to the realm of mental and relevantly similar discourse.

4. I am not here specifying necessary or sufficient conditions for "understanding an expression", nor for "using it correctly". Since these are mental terms, it is part of my view that these too are governed by criteria and so not used according to necessary and sufficient conditions. What I am doing is mentioning some of the criteria for their application.

5. The most notable contribution to his lifelong exposition of these ideas is Quine [114].

6. Since traditional epistemologists earn their living in the attempt to propose such explanations, (e.g. in the mind/behaviour case), I had better retreat to the claim that no such explanation is involved in the actual justification of ordinary assertions which contain terms governed by criteria.

7. Nothing hangs on the restriction to the present tense.

8. A person's behaviour could form part of the circumstances of danger (e.g. playing with a loaded gun); the behaviour I am
discussing is that which gives us grounds for a danger-attribution because it manifests the subject's belief that s/he is in danger.

9. On the view expressed here, it is not epistemic, or it is of a different type from that of other epistemic "because"s.

10. On the theoreticalist account, it is very hard to say what "understanding a mental explanation" amounts to. This is because, on most forms of the account, mental theory consists in lists of statements of the form: "A causes B", with no model of the causal processes involved. Hence it is rather difficult to say what I have gained when I learn that X's action was caused by the belief that p. My view allows for the understanding of mental explanation by representing it as something very different from the understanding of explanatory theories.

11. Sam Coval has suggested that the entailment is obtained if we add to the observational results some premiss underscoring the normality of the particular case, and thus ruling out bizarre counter-examples to the entailment-claim. I am not convinced that such a premiss can be stated in a way which is at once non-circular and successful in producing the required entailment.

12. I have argued elsewhere [28] that this fact does not imply scepticism about other minds and also that neither of these views is successful as a prophylactic against such scepticism; in particular, the criterial view, as it appears in Wittgenstein, was never intended to support an argument against scepticism. Such an argument cannot appeal to an account of the meaning of ordinary mental language, since it is up for grabs whether there is such a thing. Wittgenstein's strategy against scepticism is always to attempt to show that as a philosophical position it is self-defeating.

13. In the main body of the thesis, Chapter 5, section (c), I have dealt with some of the eliminative materialist attempts to equivocate on this conclusion.