UNIVERSITY OF SOUTHAMPTON

The Idea of a Cognitive Science

by
Martin Samuel Durham Ladbury

Doctor of Philosophy

Philosophy

May, 2000

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF ARTS

PHILOSOPHY

Doctor of Philosophy

THE IDEA OF A COGNITIVE SCIENCE

by Martin Samuel Durham Ladbury

This thesis offers an exposition followed by a critique of the science of cognition. As a philosophical work the concern is with the theoretical assumptions made by philosophers of cognitive science, rather than with experimental data accumulated by researchers in the field.

The exposition divides into three chapters, each devoted to a particular approach to the task of explaining intelligent behaviour. The three approaches are Computationalism, Eliminativism, and Connectionism, and the main theoretical assumptions of each are uncovered through exegesis of the work of influential advocates of the approach. The first two approaches are distinguished by their stance on the issue of the ontological status of cognitive states. Computationalists are cast as Realists about these states whilst Eliminativists are presented as Irrealists. The Connectionists are portrayed either as adopting one of these stances, or as attempting to establish a hybrid position.

The first three chapters of the critique explore arguments against Realism and, in particular, the theoretical assumption of internal content bearing states, or representations. The three lines of argument prosecuted are intended to disclose, firstly, the threat of an infinite regress of representations and representation using subjects, secondly, an impossibility of a naturalistic account of the representational content of propositional attitude states and, thirdly, an impossibility of a similar account of the normative aspect of language use and reasoning. The fourth chapter of critique explores ramifications of the arguments for Realists and for those who would decline Realism but retain aspirations of a science of cognition. The last chapter of the thesis offers a refutation of the Eliminativist version of Irrealism by undermining two assumptions required for the first of the two premisses of its argument.

The conclusion, that the idea of a cognitive science is a misconception, suggests that the route much philosophy of mind has taken, in the last quarter of a century, is misdirected.

CONTENTS

Page

Page

PREFACE

The object of this work is to provide an exposition and critique of the science of cognition. The first three chapters are devoted to general exposition and the following five present the critique. Since my thesis is philosophical in intent I have little to say about the experimental data accumulated by cognitive scientists but concentrate, instead, upon the theoretical assumptions required for the contention that cognitive activity is a proper subject matter for scientific study. Therefore, my concern is with what is said, regarding these assumptions, by the philosophers who engage in theorizing about cognition. My own contention is that, since the theoretical assumptions violate sense, experimental data will not offer a defence against the various charges of conceptual confusion I level against cognitive science in the critique. The conclusion I would press is that conceiving of cognitive concepts as apt for elucidation by an empirical science indicates a deficiency in philosophical understanding.

Admittedly, it could be said that such a bold conclusion is depreciated by my failure to provide a definitive characterization of cognitive science. There is some justification for saying this, for I cannot claim to have provided anything like an exhaustive exposition of the field. For this reason it is possible that there are those who consider themselves to be cognitive scientists, or philosophers thereof, whose positions are not touched by the arguments and objections I offer. However, since much of the critique, more or less explicitly, undermines the assumption that there are such things as cognitive states and processes, and it seems reasonable to suppose that these are the proper explananda of a science of cognition, the possible positions just mentioned would have to be considered as, at best, idiosyncratic.

In denying that there are cognitive states and processes I am not suggesting that verbs like 'believe', 'think', and 'understand' belong to a vocabulary of fictions, or that people do not, in fact, have beliefs or thoughts, or understand things. Rather, what I suggest is that to understand these terms as referential, in the manner required by an empirical science, is to misunderstand how they are ordinarily used. In basing theories on this misunderstanding the cognitive scientist cannot avoid either conceptual transgression or committing the fallacy of *ignoratio elenchi*. If cognitive scientists insist on calling computational and neurophysiological states 'beliefs' and proceed to argue that they can, therefore, discover the basic principles governing their interactions with states of a similar kind and behaviour, then they commit the fallacy as soon as they claim that their use of 'belief' is the same as the ordinary use. Of course, this is just what they do tend to claim.

As I admitted, the exposition I offer is not exhaustive. My hope is to provide a view of the theoretical landscape by mapping three main approaches to the science of cognition. The strategies I adopt are to select and examine the key works of the more influential exponents of each approach, and to cover the debates arising between those exponents. Thus, chapter one deals with Classical Computationalism, firstly, by tracing the recent philosophical background for the approach and, secondly, by examining the seminal work of J.A. Fodor, the most consistent advocate of this approach.

Chapter two takes as its subject matter Eliminative Materialism. The Eliminativist thesis admits varying degrees of severity and I present three of these by discussing the views of Paul and Patricia Churchland, and of Stephen Stich. Eliminativism of any form is very much at odds with Classical Computationalism in its vision of the form a science of cognition should take. For the former, but not the latter, psychological terms should be entirely eliminated from the language framing the laws governing the ætiology of behaviour. For our purposes, however, what is of interest is the convergence of these approaches on the position from which psychological terms are viewed as being part of a theoretical vocabulary.

At first sight it would seem that between them Classical Computationalism and Eliminativism have expended the possible stances to be taken regarding the ontological status of psychological phenomena. The first stance is that of the propositional attitude Realist, who takes 'belief', and related psychological terms, to refer to internal states, while the second is that of the Irrealist who, though agreeing that terms like 'belief' are *intended* to refer to internal states, argues that the terms lack referents. The examination of Connectionism, the third approach to cognitive theorizing, in chapter three seems to confirm this. Connectionists either tend towards the Eliminativist thesis, or they attempt to fall in with the Classicists in arguing for the existence of internal psychological states and processes. Those, like Paul Smolensky, who exhibit the latter tendency wish to distinguish their position from that of Classical Computationalism by pointing to architectural differences in their models of cognition. However, the maintenance of this distinction is problematical. I close the chapter by considering a Connectionist position which seems to fall between Realism and Irrealism but encounters difficulties in doing so.

Chapter four marks the beginning of the critique. Here my first goal is to explain why propositional attitude Realism requires a Representational Theory of Mind. My second goal is to show why the claim that there are internal representations must be supported by an account of them which does not make use of psychological concepts. To this end I invoke a Rylean infinite regress argument and try to show how the Realist response must be a retreat, from the psychological idiom, to a naturalistic idiom which makes appeal to causation in endowing internal representations with an explanatory role.

The infinite regress argument, in preventing the Realist from claiming that we have psychological attitudes to internal representations, calls into question the claim that internal states have representational content. Hence the Realist must offer a rationale for attributing internal states with meaning. In chapter five I examine causal theories of representational content and in doing so introduce, what the theorists often call, the 'Normativity Requirement'; the requirement that internal representations be capable of misrepresentation. I argue that the most plausible means to meeting the requirement is maintain, as Fodor does, that the internal representational system has a compositional syntax and semantics. I then offer reasons, some of which have a Fregean heritage, to think that the compositionalist thesis, holding that there are internal context-independent meaning elements, must be rejected. I suggest that the notion of an internal symbol must also be rejected, but present a thesis, drawn from the field of Conceptual Role Semantics, which argues that internal states symbolize in virtue of their causal role. This argument serves to connect the point at which we ended chapter four with the next phase of the critique.

The claim, that internal states have meaning in virtue of their causal role, can be upheld only if normative evaluations of uses of words are applicable to the products of mechanistic processing. Similarly, the claim, considered at the end of chapter four, that internal representations contribute to rational behaviour because of their causal properties, is intelligible only if the rationality of behaviour (its compliance with standards of rationality) can be described drawing only from the resources of causal explanation. I argue, in chapter six, that language use and reasoning cannot be the product of 'cognitive' mechanisms because these activities allow of normative evaluations of the kind appropriate to followers of rules, and mechanisms do not follow rules.

In chapter seven I try to show how much damage is inflicted, upon the idea of a cognitive science, by the course of the argument of the three preceding chapters. The contention that internal processing can account for linguistic content and cognition would require both that there exists a framework of principles governing the processing, and that those principles offer standards upon which normative judgement can be based. In other words what is required is a normative framework which allows internal states, and the processes ranging over them, to take on meaning and support cognitive states. Not only do I argue, by citing some remarks of Wittgenstein, that a normative framework of fixed rules fails to determine meanings, but I also reiterate the point that internal states of physical system could not follow rules. Consequently we must reject the notion of an internal representational system—an internal language—as incoherent. In the remainder of the chapter I try to show that the Connectionist approach to cognition can offer no retreat for the Realist and that, consequently, it might appear that the remaining explanandum for a science of cognition would be linguistic competence. I add further weight to the arguments, of the last three chapters, to push to the conclusion that linguistic competence is not to be accounted for by looking within language users.

The eighth, and final chapter, revisits the thesis of Eliminative Materialism. In presenting the thesis we find that, despite appearances, the Eliminativist and Realist share common assumptions about the nature of psychological language, *viz.* that it is referential and aspires to nomic generalization. In criticizing both these assumptions I undermine the basic premiss upon which the Eliminativist conclusion is supposed to follow; the premiss that psychological language is theoretical. The topic of normativity re-emerges, firstly, when I suggest that if 'folk psychology' was a theory then there would be no normative evaluation of reasoning and, secondly, when I observe that without psychological concepts we could not distinguish sense from nonsense.

Believing, understanding, remembering, or thinking that *p* are not internal states and processes. The Realist cannot present a plausible account of how an internal state can have propositional content or play a part in explaining action, and the Eliminativist cannot justify the claim that in ascribing beliefs, and the like, we are committed to the assumption that they are internal states. The latter, in suggesting that there are no such things as beliefs and that we should, therefore, explain behaviour without appealing to them, indulges in the same misuse of language as the former. Not only are concepts such as 'representation', 'language', and 'concept' distorted, but also the cognitive concepts populating the very domain of explanation are entirely misconceived. I find it difficult to see how this position can be rectified in such a way that we can retain any semblance of a cognitive science.

I am very grateful to Professor Anthony Palmer for supervising the production of this thesis. He kindly drew attention to, and tried to remedy, my own philosophical deficiencies and, more than once, pointed the way to lines of argument more fruitful than the ones I had proposed to take. Of course, he is not to blame for any ill-chosen or inadequately constructed lines of argument, nor for any mistakes in my reasoning.

I should also like to express gratitude to my wife, Victoria Ladbury, without whose encouragement and support I would have been unable to embark upon, and continue, the course of research of which this thesis is the culmination. My colleagues Geoffrey Eavy, Graham Stevens, and Vincent Iyengar, have offered moral support and have continually shown willingness to discuss, and offer suggestions regarding, my work, and to them I am also indebted. A similar debt is owed to various members of the Department of Philosophy who have listened to, and offered criticism of, parts of the thesis. Lastly, I should like to

thank the School of Research and Graduate Studies and the Department of Philosophy (both at the University of Southampton) for the award of a two-year Teaching Assistantship.

CHAPTER ONE

COMPUTATIONALISM

1 TURING AND THE PHILOSOPHY OF MIND

One of the earliest versions of the Computational Theory of Mind was offered by Alan Turing in his 1950 article ‘Computing Machinery and Intelligence’. It is convenient to view the history of the philosophy of mind as a linear succession from Dualism to Behaviourism, from here to Central State Identity Theories, and from these to the Computational Theory of Mind. The view is convenient because it affords a picture of steady progression, for each new conception of mind seems to be more tenable than its predecessor. Thus Behaviourism did away with the dubious ‘Ghost in the Machine’ of Cartesian Dualism, Identity Theories corrected Behaviourism by placing agency and intentionality within the body, and the Computational Theory of Mind (which started life as Turing Machine Functionalism), avoided the standard objections to identifying what is individuated by means of introspection (a sensation, for instance) with what is objectively individuated (a brain process).

It might be argued that the chronological facts do not bear this historical thesis. J.B. Watson's ‘Behaviourism’ was published in 1924, some quarter of a century before Gilbert Ryle's consummate exorcism of the Cartesian ‘Ghost’ in *The Concept of Mind.* Turing's aforementioned article predates J.J.C. Smart's ‘Sensations and Brain Processes’ by nine years and even B.F. Skinner's ‘Science and Human Behaviour’ by three. If we take these as seminal books and articles of the last three, chronologically ordered, of the four philosophies of mind listed above, then the progressive historical perspective seems inconsistent with the facts.

However, by way of partial reply to this challenge, there is an argument to the effect that the Computational Theory of Mind did emanate from and improve upon Behaviourism, for Turing's article could be viewed as implying a behavioural account of cognitive processes.

The question Turing addresses in 'Computing Machinery and Intelligence' is 'can machines think ?', and the criterion he suggests for determining the answer is whether a digital computer of unlimited capacity could win at the Imitation Game. The game would be played in the following way: A digital computer, A, and a person, B, would remain in one room whilst another person, C, would be in another. C would be told that the first room is occupied by a person and a computer and C's task is to ask A and B questions on any matter and judge from their answers which is the person and which the computer. The manner of communication must not be helpful to C so type-written questions and answers would be used. C wins if she guesses correctly whilst the computer wins if she does not. B can help C as much as he can, but since A can type things like 'Don't listen to him, I am the person', the effectiveness of B's help will be negligible.

If the computer wins the game and is judged to be a person, then the conclusion one might draw is that it can think. Turing shies away from stating this consequent because he believes the original question 'can machines think ?' to be 'too meaningless to deserve discussion' preferring, instead, to ask whether it could pass his test (that is, win at the game). His confidence, that a digital computer could, led him to proclaim;

> 'I believe that at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted.' (Turing, 1950, p.142).

For Turing's test to be of real significance to the philosophy of mind it must be that passing the test is indicative of the possession of cognitive capacities. The argument might go that since, in the test, the criterion for deciding whether a subject has such capacities is the linguistic behaviour it displays, the capacities themselves are wholly explicable in terms of such behaviour. This reasoning is echoed in Turing's dismissal of the 'Argument from Consciousness' which maintains that machines lack subjective experience. Turing says that such an argument can be made to collapse into Solipsism and reiterates his confidence in his Imitation Game, and other such performance tests, as the appropriate means for assigning intelligence to systems (ibid., pp.145-147).

Although there are hints of Behaviourism in Turing's article it is also clear that he wants to draw analogies between machines and the human brain, and these suggest an internalistic account of cognition.[1] Indeed, what is interesting is that Turing's modelling of his test on the imitation game prohibits use of the test to validate a behavioural account of intelligence.[2] The original imitation game, according to Turing, is one in which A is a man, B is a woman and C has to guess which is which whilst being hindered by A and helped by B. If A succeeds C will identify him as the woman. This is the corollary of the computer (A) being identified as a human. If, in the Turing test, C's identification of A as a thinking being, on the strength of A's behaviour, warrants the conclusion that A *is* a thinking being, then, by analogy, C's identification of A as a woman, in the imitation game, should warrant the conclusion that A *is* a woman and this is plainly wrong. A more correct analysis would be that, in the imitation game, A's ability to mislead C is owed to the fact that he can think like a woman, for this would allow us to conclude that the reason why a computer might pass the Turing test is that it can think like a human and, hence, that it can think. On this analysis it is not the behaviour of

[1] Turing believes that a machine capable of winning the Imitation Game could be created by educating the machine in a way analogous to that in which we educate children. To this end he proposes the designing of a programme to simulate a child's mind. See also 'Intelligent Machinery', pp.120 & 121, in which Turing likens the infant human cortex to an unorganized machine (Turing, 1948.). From the Behaviourist's point of view the input and output of the computer must correspond to the stimuli and responses of the human if the former is to be attributed with the cognitive capacities of the latter. By emphasizing analogies between the mediating processes of each, Turing might be taken to be suggesting that thinking is not simply a form of behaviour, but is in fact the process which produces that behaviour.

[2] I have taken this observation and the subsequent analysis from D. Proudfoot, 1997, pp.191-194.

the computer alone that allows it to pass the test, but the thinking which informs the behaviour. Thus, if Turing did have a behaviouristic construal of his test, it was misconstrued.

The interesting point that emerges from this discussion is that Turing lays the footings for, but does not fully erect, the Computational Theory of Mind. He predicts that machines will be able to perform in at least some of the ways humans perform when they think, thus qualifying them for the epithet 'thinking beings'. However, he does not formulate the subjunctive which encapsulates the rationale behind the Computational Theory of Mind, *viz*., 'If we could create a computer which thinks, then we would have shown that thinking is a computational process'.

The subjunctive is clearly anathema to a purist Behaviourism because of the implication that thinking is a process internal to that which thinks. That is, the Behaviourist might happily accept the antecedent but must reject the consequent. I will not attempt to settle the question of whether Turing did not speculate about what happens in humans when they think because he believed 'thinking' describes a kind of behaviour. What is significant for the purpose of this chapter is that he paved the way for a computational account of cognition.

To return to the discussion of historical progression in the philosophy of mind, we can see that if Turing is attributed with a Behaviourist's understanding of cognitive capacities, then the thesis that the order of succession was from Dualism to Behaviourism; from here to Identity Theories; and from these to the Computational Theory of Mind, is not contravened by the date of Turing's article. Furthermore, given this attribution to Turing, the two oft cited typical articles of the Central State Identity Theory fit quite neatly into their chronological slot. Smart's 'Sensations and Brain Processes' was published in 1959, while 'Is Consciousness a Brain Process ?' by U.T. Place appeared in 1956.

Unlike the Behaviourists, Smart and Place believed that in order to understand the nature of mental episodes we should concentrate our endeavours not on the study of behaviour and its surroundings, but on investigating the activity of the brain or, more properly, the Central Nervous System (CNS). The reason why we should do so, they thought, is because mental episodes just are episodes of the CNS.

It is worthy of note that the materialism of Smart and Place did not entirely turn its back on Behaviourism. Place suggested that propositional attitudes should be analysed in terms of dispositions to behave (Place, 1956. p.106.) whilst Smart professed sympathy with 'expressive' accounts of sensation statements, though he did not think they would 'quite do the trick' when it came to explaining what such statements are about (Smart, 1959. p.119.).[3] However, Place believed, behavioural accounts of propositional attitudes notwithstanding, that;

> 'there would seem to be an intractable residue of concepts clustering around the notions of consciousness, experience, sensation, and mental imagery, where some sort of inner process story is unavoidable.' (Place, 1956. pp.106 & 107.).

The view was echoed by Smart (see the first paragraph of Smart, 1959).

Smart and Place, then, were proposing an identification of brain processes with, what might be called, the qualitative content of experience or, what Wilfred Sellars called, 'raw feels' (Sellars, 1965). Their thesis, despite its appealing simplicity and perhaps because of it, faced insurmountable objections. Many of these centred around the fact that the means by which we individuate sensations or conscious episodes are entirely different from those by which we individuate brain processes. This is apparent when the means of individuation are observational, since conscious episodes are supposedly discerned through (necessarily subjective) introspection while brain processes can be (objectively) measured. No amount of inward looking by the subject will give him or her a glimpse of the brain process identical to,

[3] The 'expressive' account Smart is refering to is that of Ludwig Wittgenstein (as exemplified in §244 of the *Philosophical Investigations*, Wittgenstein, 1953.) where he suggests that self-ascriptions of pain are to be understood, sometimes, as expressions (in the way that crying is an expression) rather than as reports on a state of affairs. Smart (ibid.) was casting Wittgenstein as a Behaviourist in his discussion and it should not go unmentioned that many, including myself, would disagree with such a charaterisation.

say, a yellowish-green after-image, and no measurement or observation of the brain process will reveal the yellowish-greenness of the after-image.

The objection more commonly takes, and is greatly improved by taking, a linguistic form whereby the means of individuation is assumed to be descriptive rather than experiential. In this form the thrust of the argument is that the predicates used to describe brain processes, on the one hand, and conscious episodes, on the other, are not and could not be intersubstituted as ought to be possible if they describe the same event. Thus, the after-image can be described as 'yellowish-green' but the brain process cannot, and the brain process can be described as consisting in the transfer of transmitter molecules but the after-image cannot .[4]

It would be wrong to suggest that Smart and Place, or many of the others who had adopted the Identity Thesis (such as Feigl and Rorty), had not anticipated the main objections. Indeed, Smart's paper is largely made up of objections, of the form I have given, and his replies to these. Such awareness can be attributed to Place (Feigl and Rorty) also. However, the methods of dealing with the objections were themselves problematic and the Identity Theory, in this form, lost favour.[5]

Before moving on from the Identity Theory I should like to draw attention to one of the counter arguments fielded against the individuation objection in its linguistic form. It is offered by Richard Rorty in his article 'Mind-body Identity, Privacy, and Categories' (Rorty, 1965.). In his view we should not phrase our psycho-physical identity statement in the form,

[4] For a discussion of the linguistic form of the individuation objection see Sellars, 1965.

[5] By 'this form' I mean the Type Identity Thesis which identifies a type of mental episode with a type of physical event. A major impetus towards rejecting this thesis was the consideration that a creature with a very different physiological constitution to ourselves, such as a Martian, could be described as experiencing the same types of conscious episodes as we do when clearly its neural events would be of a different type (assuming it has a CNS, of course). The Token Identity Thesis, which identifies a token of a mental episode type with a token of a neural episode type, had Donald Davidson as its earliest advocate. He presented his thesis of Anomalous Monism in his paper ' Mental Events' (Davidson, 1970) and continued to support it a quarter of a century later (see Davidson, 1993). On this thesis, like that of the Causal Theorists and the Functionalists, mental episodes can be identified with physical ones by virtue of their causal roles, rather than their qualitative contents, thus allowing psycho-physical identifications to be intimated without vulnerabilty to objections arising from the issue of qualitative content.

‘Sensations are identical with certain brain processes’, but should, instead, assert that ‘What people now call “sensations” are identical with certain brain processes’. By doing so we are avoiding the ‘category mistake’ of mixing mental and physical descriptions as applied, theoretically, to the same event. As Rorty explains, ‘there is no reason why “what people call ‘x’” should be in the same “category” (in the Rylean sense) as “x”.’ (Rorty, 1965. p.19). In other words, since ‘sensations’ in the new assertion is being mentioned rather than used, what the quoted term refers to is brought into the asserted equation without its mentalistic baggage.

I do not wish to comment on the plausibility of Rorty's view, but include it in this historical survey only to introduce one branch of a bifurcation in the materialistic approach to the philosophy of mind. For what Rorty was envisaging, implicitly, is a future in which mentalistic expressions such as ‘sensation’, ‘feeling’, and even ‘pain’, or ‘ache’ will be obsolete. Our use of psychological terms will be superseded when a neuroscientific vocabulary is in place. This is the position of the Eliminative Materialist, and it now an established position within cognitive science (and one we shall explore more fully in the next chapter).

I said that Rorty's article marks a bifurcation one branch of which leads to the Eliminative position, and will now, by exploring the other branch, return to the theme of this chapter. One path to maintaining a materialistic ontology whilst avoiding the problems of the psycho-physical Identity Theory was simply to remove the psychological from the equation, as Rorty proposed. Another was to make more sophisticated the characterisation of psychological states, that is, to replace the characterisation which appeals to qualitative content with one appealing to functional role. This was the path taken by the Functionalists.

2 PUTNAM AND THE MIND-BODY PROBLEM

As I said above, Turing's 'Computing Machinery and Intelligence' paved the way for the Computational Theory of Mind by posing the question 'can machines think ?', but did not venture the subjunctive 'If so, then thinking is a computational. Perhaps this is not surprising since the question is about machines whilst the subjunctive makes a statement about the nature of thought, and Turing's interest was in the potentials of computing machinery as opposed to speculative psychology, or philosophy of mind. Had Turing's primary concern been the mind-body problem he might well have been less mechanico-centric in his speculations. Be that as it may, a correlation of computers with minds certainly was mooted a decade after Turing's article.

In 1960 Hilary Putnam published a paper entitled 'Minds and Machines' (Putnam, 1960) in which he suggested that the traditional question as to the relationship between mind and body, with all its attendant conceptual considerations, would be mirrored by (because it is logically analogous to) the question of the relationship between a Turing machine's functional states and their physical realisations. For example, an Identity Theorist like Smart might have supported the proposition;

1) I am in pain if, and only if, my c-fibres are being stimulated.

This proposition has its Turing machine analogue in the following proposition which the machine could 'assert';

2) I am in state A if, and only if, flip-flop 36 is on.

It becomes apparent that 1) and 2) are logically analogous on considering that both are synthetic propositions; for it would not be contradictory to postulate cases in which the antecedent condition holds when the consequent does not and vice versa. Furthermore, objections which may be levelled at 1) can be transposed, in the same form, into objections to

2). Thus the objection that pain and c-fibre stimulation can not be connected by a biconditional, because they denote occurrences individuated by entirely different means, is paralleled in the case of the Turing machine's state and its physical configuration. The machine's individuation of its state A and the Identity Theorist's of his pain are not achieved via inferences from observations while the flip-flop switchings and c-fibre stimulations are.[6]

Putnam's conclusion was that just as the question of the 'identity' or 'non-identity' of a Turing machine's logical and structural states is of little general importance, so the analogue of this problem for human mental and physical states is equally unimportant (ibid., p.384). Why Putnam felt that we should be satisfied with this conclusion will become clear when we look more closely at his thesis.

So, in 1960 Putnam constructed an analogy between the conceptual considerations of identifying people's mental states with their physical states on the one hand, and machines' logical with their structural states on the other. Soon after, in his paper 'Some Issues in the Theory of Grammar' of 1961, he suggested that 'there are many considerations which point to the idea that a Turing machine plus random elements is a reasonable model for the human brain.' (Putnam 1961, p.102).[7] Thus the analogy is broadened from having as its analogues identifications (and objections to these) of state types within two systems (human and mechanical), to having the systems themselves as analogues. In this context the broadening occurs in order to take in certain linguistic capacities (see note 7).

[6] As I have explained, the Identity Theorists held that we individuate sensations introspectively, as 'raw feels', and it is this that provides the objection with its basis. Putnam prudently suggests that not all first-person avowals of sensations are based on introspection (after all, it keeps his machine-body analogy afloat by avoiding the snag of having to maintain there are introspective machines)(See Putnam, 1960. p.368). Indeed Functionalists, unlike the Identity Theorists, need not rely on introspective evidence as the means of individuation of a sensation. The sensation, as referent, does not dissolve as a result (as it does for the Eliminative Materialists) because it is individuated as a functional state.

[7] The possibility of constructing a digital computer which includes a random element is mooted by Turing (Turing, 1950. p.138) where he explains that the random element, when incorporated into the machine's operations, may give it the appearance of having free will. Putnam's suggestion is made in the context of an argument that the grammatical sentences of a language are a recursive set and, therefore, that the sentences of a language may be classified as grammatical and ungrammatical by a mechanistic system. In practice human classifications are likely to be idiosyncratic, hence the need for a random element in a machine modelling human linguistic performance.

However, it is in his 'Philosophy and Our Mental Life' (Putnam, 1973) that Putnam states his case most perspicuously. He says;

> 'The concept which is key to unravelling the mysteries of the philosophy of mind, I think, is the concept of *functional isomorphism*. Two systems are functionally isomorphic if *there is a correspondence between the states of one and the states of the other that preserves functional relations.*' (ibid., p.291)

For example, suppose we take, as our systems, two computing machines in which the functional relations are sequential so that, for example, state A is always followed by state B. In this case F will be a functional isomorphism when state A is followed by state B in system 1 if, and only if, state F(A) is followed by state F(B) in system 2. Putnam gives as an example of a functional relation the print-out relation 'When π is printed on the tape, system 1 goes into state A'. If system 1 goes into state A on having π printed on its tape then system 2 must go into F(A) in the same circumstance if both are to be deemed functionally isomorphic.

It should be noted that Putnam explains the notion of a functional description in terms of a Turing machine table, 'a standard style of program', and he intimates that there should be a similar normal form of description for 'systems like ourselves' (ibid., p.292).[8] Indeed it is this latter supposition which indicates how the concept of 'functional isomorphism' is to unravel 'the mysteries of the philosophy of mind'.

If we compare systems with reference to their functional descriptions we might find we have two calculating systems, for example, which are functionally isomorphic yet physically

[8] A Turing machine table depicts how the last state of a machine, plus its input, determines its present state thus:

		Last State		
		A	B	C
Input	Y	B	C	A
	N	A	B	C

Here the input Y represents the command to change state while N gives the opposite command. If the machine was such that in state C a light bulb was lit, and this was the only visible output of the machine, the output table would be as follows:

State	A	B	C
Output	Off	Off	On

For a fuller explanation see Turing, 1950 pp.139&140, or for a more complex schema see Putnam, 1960, p.365.

very different. One may be made of electrical components while the other is made of cogs and wheels. However, since our concern is purely with the functional, the components from which the systems are made are irrelevant. Now if we believe, as Putnam does, that the notion of functional organisation applies to anything to which the notion of a psychological theory applies, then it is possible to give functional descriptions of humans such that two could be functionally isomorphic. Further, we might suppose that each of these humans occupies a possible universe, one in which people have souls operating on their bodies through pineal glands in the brain, the other in which the people have brains only. Here we have a case in which two humans are psychologically identical yet the realization of their psychological states, that is the medium in which the states are instantiated, is as ontologically disparate as we can get. Putnam's point, then, is that ontology is irrelevant to psychology when the latter is understood as applying to functional systems.

Indeed, the ontological position of the Identity Theorist, that mental episodes are identical to physical ones, cannot be correct, as Putnam explains:

> 'For it is clear from what we already know about computers etc., that whatever the program of the brain may be, it must be physically possible, though not necessarily feasible, to produce something with that same program but quite a different physical and chemical constitution.' (ibid., p.293)[9]

The physical (or otherwise) realization of a mental episode is accidental from the point of view of psychology, and plainly it is wrong to identify the episode with its realization since the same episode could be realized in any number of systems, all with differing constitutions. This is because the episode can be given a functional description.

Putnam's thesis, then, is that the explanation of human behaviour is not dependant upon whether we have immaterial souls or are purely material organisms.[10] Rather the explanation is

[9] This is the rationale for the 'Martian' argument mentioned in note 5.

[10] Putnam pledges allegiance to Aristotle in insisting that what is of interest is our intellectual form, not the matter in which that form is manifested.

to be effected through an exposition of our functional organisation, and this has nothing to do with ontology. The analogy between humans and Turing machines, which makes the latter a 'reasonable model' for the former, can be drawn because both, on Putnam's view, have functional organisation.

Putnam is at pains to establish that he is not suggesting that a human being *is* a Turing machine of sorts or that psychological states *are* Turing machine states. One good reason for denying these identifications is that Turing machine states are *total* in the sense that they wholly specify the present condition of the machine. Psychological states, in contrast, are not total in this sense because a person may be in a state of pain, for example, at the same time as he or she is hearing a shrill whine, or intending to say 'three' (ibid., p.298). Nevertheless, the analogy between humans and Turing machines is upheld by Putnam. Though he eschews any attempt at reduction from minds to physical systems, he does intimate that a physical system—a computer for instance—might have a functional description which at least approximates to a psychological theory (ibid., pp.301&302).

Accepting what Putnam said in the three cited papers we can speculate that it would be possible, in theory, to create a computer functionally isomorphic with a human mind. What is needed is a suitably coherent psychological theory which can be expressed as a description of a functional organisation; an organisation which can be replicated in the electronic circuitry of a computing machine. If this is possible then it looks as if we can say that thinking is a computational process. Not only that, but we may also reaffirm Turing's conjecture that machines will think (or at least will be spoken of as 'thinking'), and we can reaffirm it on grounds other than the behaviour of the machine, since the function of thought need not always involve external behaviour.

Now I propose to look at an account of the functional organisation of the mind which has been prevalent among the adherents to the Computational Theory of Mind.

3 FODOR AND THE LANGUAGE OF THOUGHT HYPOTHESIS

a) Background

One way to deduce the functional organisation of the mind is to consider what kind of tasks it must perform. Let us assume that the overall task of a mind is to receive stimuli and produce responses which are appropriate to these. Whether the responses are appropriate will depend on whether they conform to various standards of rationality which might be supposed to have some teleological basis. For example, given that avoidance of physical injury is a goal it is rational to pursue, a man's response of flight will be wholly appropriate when his sensory stimuli amount to the perception that an angry bull is bearing down upon him.

Unlike Behaviourism, in which the emphasis is on stimulus and response, and the appropriateness of the latter does not result from the subject's cognitive competence but from his or her conditioning, the Computational Theory of Mind posits processes as intermediaries between the input (a stimulus) and output (a response). Rationality is introduced through the processes' matching of inputs to outputs according to rules or norms. It is these processes, and the structures they range over, which constitute our mental episodes and states. Thus we find that the processes have two forms of description; one in which they are individuated by their functional roles as the processes which produce a certain output when triggered by an input (given a certain structural organisation of the system); and another in which they are individuated by the everyday mental taxonomy using terms such as 'see', 'hear', 'think', 'feel', 'hope', and 'calculate'. These and other terms make up the vocabulary of what is often called 'Folk Psychology'. It is the burden of the Computational Theory of Mind, therefore, to support the suggestion that they correspond in some way to, or can be accounted for by, the processes bearing functional descriptions.[11]

So, if the mind is a system of structures and processes designed to produce appropriate output on receipt of input, an account of its functional organisation will be an account of the structures and processes and their relations. A central task of the Computationalist enterprise is to give such an account, and an attempt to do so can be found in Jerry Fodor's book *The Language of Thought* (Fodor, 1975).

It should be observed that there were multiple routes leading to the establishment of the Computational Theory of Mind. The theory was not simply a response to unsatisfactory solutions to the mind-body problem, or an adaptation of one or more of these. Developments in disciplines other than philosophy were setting the scene for its appearance.

Clearly the arise of computer technology was a prerequisite for a theory which models the mind on the functional organisation of an, albeit idealized, computer. When one considers that the first multi-purpose electronic computer (the Electronic Numerical Integrator and Calculator) was built by John Von Neumann and his team in 1946, and yet just fifty years later almost every form of manufacture, communication, administration, and transportation now involves electronic computers, in one way or another, the advance and spread of computer technology has been on the proportions of an epidemic.

Speculative projections of human cognitive attributes onto electronic computer functions were made very soon after these machines first appeared. As early as 1950 Turing had entertained the idea of an intelligent computer, and by 1959 Marvin Minsky and John McCarthy had set up the Massachusetts Institute of Technology Artificial Intelligence Project.

[11] Whether the functional processes of the human brain can be expected to gravitate into sets which correspond to 'folk psychological' types (such as belief, hatred, or expectation) is a matter over which Computationalists have held conflicting opinions. Jerry Fodor, for his part, can best be understood as a Realist about such psychological types. That is, he believes that the brain's functional organisation reflects and validates our folk psychological taxonomy. See later chapters for futher discussions of this matter.

If what I have said above about Turing's 1950 article is correct, the question of whether an intelligent computer can be produced is conceptually separable from that of whether intelligence is a computational phenomenon. At least it is separable provided we accept the behavioural criteria of intelligence presupposed in the Turing Test. Thus, although the likes of Turing and his contemporaries in the field of Artificial Intelligence (hereafter, AI) were not in the business of speculative psychology (not officially, at any rate) they were clearly doing a great deal to help development of computational models onto which future psychologists could map cognitive processes.[12]

Given the title of Fodor's aforementioned book it is not difficult to see that developments in the field of Linguistics played a part in the formulation of subsequent theories of cognition. Linguistics underwent a metamorphosis at the hands of Naom Chomsky who, in the late fifties, help shift the emphasis of this discipline from study of the structures of natural languages to the deduction of the structures which must be in place before languages can be learned. The enterprise had been prompted by the perceived need for an explanation of how a language, with a potentially infinite number of legitimate locutions, can be learned on the basis of exposure to a finite number of examples. A competent speaker of English, for example, can construct and understand grammatical sentences which he or she has never encountered before, and for Chomsky this fact had to be explained by revealing the nature of the human organism rather than the nature of the language itself.

The key to the explanation Chomsky advanced is the grammatical aspect of language. Broadly speaking it is the grammar of a language which legislates between permissible and

[12] A distinction was made between 'strong' and 'weak' AI by John Searle (see Searle, 1980). It is useful as a means of removing ambiguity as to the status of computational models of intelligence. Such models can be seen as mere simulators of, for example, human linguistic competence, in which case they are weak or, alternatively, the models can be intended as replicators of human linguistic competence so that the processes described in the computational model are to be understood as having the same functional role and description as those occurrent in human brains. The latter would be the view from the strong stance. Turing is probably best understood as presenting a weak AI thesis. His intelligent machine would simulate human linguistic competence by performing as we do, but the processes which are supposed to constitute the human competence and those which constitute that of the machine need not be alike in any way for the simulation to be effected. (However, see Turing, 1950, p.156 for hints of a stronger AI approach.)

non-permissible utterances, thus allowing those who produce predominantly the former type to be accredited competence in the language. A grammar can be expressed as a series of formal rules governing the structure of sentences. Now, what is significant about such a grammar is that, provided that at least some of its rules are recursive, it can be finite yet capable of legislating over an indefinite number of sentences. The competence of a speaker of a language could thus be accounted for by his or her knowledge, or possession, of the grammar of that language.[13] Knowledge of the formal system would facilitate inferences and generalizations to novel, but legitimate, sentences from the limited set of sentence exemplars encountered by the speaker.

In the late sixties Chomsky augmented this explanation of linguistic competence with the supposition that an innate knowledge of grammar is possessed by all language speakers which allows them to learn to use an indefinite number of sentences.[14] Based on this assumption, Chomsky's task was to produce a theory of grammar which would describe the structures which were innate in language speakers.

My account of the basic ideas behind the theory of Generative Grammar, first presented by Chomsky in *Syntactic Structures* (Chomsky, 1957), is simplified and selective for I have said nothing of considerations of syntax, semantics, or phonology, for example. However, it does, I hope, serve the purpose of suggesting how Linguistics, in the hands of Chomsky in the late fifties, extended its borders beyond the study of natural languages into the realms of speculation about the psychological requirements of language speaking. Thus the space was created for occupation by the new field of Psycho-linguistics. It is from this field that Fodor

[13] Note that this grammar would be descriptive of the actual rules employed by speakers of the language rather than prescriptive of what rules should be followed. Some idealization must occur in specifying the rules of a language in order to avoid the risk of an unmanageable grammar which allows for idiosyncrasies in a speaker's linguistic performance. See note 7 for a similar point, made by Putnam, relating to the construction of a digital computer which would model human linguistic performance. The applicability of Chomsky's linguistic theories for computational models of mind (in Putnam's case simulating, rather than replicating, models) was appreciated well before Fodor's definitive statement of the Computational Theory of Mind.

[14] Chomsky believed that all natural languages could be governed by the rules of a universal grammar. If this were not so then there would be a problem of explaining the fortuitousness of the presence of, say, an innate grammar for Portuguese in a child born in Portugal, as well as a related problem of explaining multi-lingual competences.

draws much of the empirical support for his philosophical arguments (as can be seen in Fodor, 1975, particularly Chapter 3) and it is, perhaps, because of this that he classes himself as a speculative psychologist in the preface of *The Language of Thought.*

The marriage of psychology and linguistics was not just a result of advances in the latter, for the belief in mental structures was being promulgated by psychologists also. Jean Piaget, for instance, in his attempt to explain how intelligence develops in humans, posited a set of structures, or 'schemas' which organise experience (Piaget, 1971). Piaget suggested that as a person grows from birth to adulthood these schemas are formed and adapted to provide him or her with the appropriate conceptual framework for understanding and acting upon his or her environment. Identifying four stages of development through which this conceptual framework is formed Piaget held that at each stage the schemas, when compounded, may be seen as constituting a logic. The assimilation of new schemas is seen as progressive so that the child's development entails the learning of a series of logics with increasingly more power for handling external stimuli.[15]

Knowledge and intelligence, for Piaget, are thus gained through a process of structural development in the human mind. These structures facilitate, and determine the nature of, cognitive processes. That is, they must be appealed to in explaining human perception, understanding, thinking, memory, problem-solving, and so on.

I confine myself to the work of Piaget in exemplifying the steps taken in psychology just prior to Fodor's publication of the *Language of Thought* firstly because Fodor himself discusses it at some length, and secondly because it was so influential in the field. When combined with what I have said about Chomsky's theorising this barest of sketches of Piaget's

[15] Strictly speaking it is not until the third stage of development, between the ages of around seven to ten, that the child uses schemas constituting a logic *per se*. According to Piaget it is only at this stage that the child begins to employ reversible operations, such as ordering objects according to size (which is reversible only if the child understands that the ordering depicts increasing as well as decreasing size), and principles of quantitative conservation (as exhibited when a child understands that, say, a quantity of fluid poured from a short wide container into a tall thin one is still the same quantity). Prior to this the child can, at best, be said to be at "semi-logical" stages (see Fodor, 1975, pp.88&89).

concerns serves to give some idea of the intellectual environment from which Fodor's account of the mind sprang.

This section began with the reasoning that an account of the functional organisation of the mind will entail a description of the structures and processes which constitute mental activity. Piaget was concerned with the development of cognitive, or knowledge related, faculties and saw developed structures as forming a logical system for dealing with experiences. Chomsky, on the other hand, was attempting to explain linguistic abilities, or competencies, and so the mental structures he described were, supposedly, grammars. The thesis Fodor presented in *The Language of Thought* had the potential to explain not only cognitive processes and linguistic competence, but another aspect of psychological discourse also; namely propositional attitudes. Like Chomsky and Piaget, Fodor also postulated a structural system and it was this that he termed the 'language of thought' (henceforth abbreviated to the LOT).

b) The Hypothesis

To continue the exposition along the lines suggested in the last section's first sentence I will sketch Fodor's LOT hypothesis by demonstrating how it can be used to explain a cognitive task. Understanding a proposition may be used as an example of a cognitive process, a propositional attitude, and a sign of linguistic competence so I will take this as the explanandum.

According to Fodor's model, in understanding the utterance 'It is raining' the hearer translates a 'wave form' into the 'message' that the speaker wishes to communicate. In Fodor's words:

> 'A speaker is, above all, someone with something he intends to communicate. For want of a better term I shall call what he has in mind a message. If he is to communicate by using a language, his problem is to construct a wave form which is a

> token of the (or a) type standardly used for expressing that message in that language.'(Fodor, 1975. p.106)

The wave form expresses the message because it can be 'mapped' from it and, again in Fodor's words;

> 'The character of each mapping is determined, *inter alia*, by the conventions of the language the speaker and hearer share. Verbal communication is possible because the speaker and hearer both know what the conventions are and how to use them: What the speaker knows allows him to pick the value of U [the utterance] which encodes a given value of M [the message], and what the hearer knows allows him to pick the value of M which is encoded by a given value of U. The exercise of their knowledge thus effects a certain correspondence between the mental states of the speaker and hearer:' (ibid., p.108)

So, when it is raining and Noah wants Nelly to know this, he has a message in mind which he intends to communicate. Noah proceeds by mapping the message onto the appropriate wave form which, convention dictates, will convey this message. Because Noah and Nelly have learned English the wave form is the spoken sentence 'It is raining'. On hearing the wave form Nelly applies her knowledge of her natural language, English, to use the sentence to produce in her mind the message Noah intended to convey. When Nelly has got the message, so to speak, she has understood Noah's sentence 'It is raining' and she and Noah are in corresponding mental states.

This description of communication is not difficult to grasp and is, at first sight, an appealing one. From such a view point there would appear to be little more to verbal communication than there is to 'communication' between Facsimile machines, for example; the principles are the same. The once, seemingly, nebulous phenomenon of understanding is rendered as a process by which one artefact, a wave form, is simply transformed into another, a message. Processes of this kind are well known (though not necessarily in detail) by all of us who use tele-communications of one sort or another. Of course, the analogy is not quite right as a Fax machine would not be attributed with the understanding of the messages it reproduces (hence my use of quotation marks around the word 'communication' above) whilst, for humans, the reproduction *is* the understanding of the message, or so it seems on Fodor's account. The pertinent difference at this stage would be that Fax machines lack knowledge of the natural language in which the messages are phrased. Be that as it may, Fodor's explanation utilizes a model for which we already have concrete examples. I would suggest that it is this factor which makes the Computational Theory of Mind generally appealing. Fodor's account of communication is dense with assumptions so I shall extend my exposition by revealing some of these and providing Fodor's rationale for them.

To begin with, the process of understanding by transposing wave forms into messages is not supposed to be something a hearer is aware of. That is to say, though Nelly understands Noah's utterance, the process by which she understands is not a conscious one. To maintain the contrary would be perverse since we can all testify that we are not conscious of having to transform the sentences we hear into messages (or anything at all for that matter).

A consequence of the assumption that the process of understanding involves translation is that the spoken language is understood only indirectly. For the English sentence Noah utters is understood by Nelly only via the message he conveys to her which is itself expressed in the LOT. So it is assumed not only that the process of understanding is unconscious, but also that it requires a LOT. An account of the LOT is now owing.

Fodor is at pains to explain that the LOT is not a natural language in that it is not public and learned. His argument in support of this may be summarised as follows:

a) People acquire competence in natural languages.

b) A person's acquisition of competence in a natural language entails he or she has a theory thereof because learning a language is a matter of making and confirming hypotheses about the truth conditions associated with its predicates. An example of such a hypothesis might be;

1) [P*y*] is true iff G*x*.

c) This requires that there be a language in which to formulate such hypotheses and, in the case of 1), G will be a predicate of that language coextensive with the predicate, P, of the natural language.

d) Clearly this second language can not be learned because, by b) and c), this would require a third language, which would require a fourth, and so on.

e) Since we do learn languages we must conclude that the LOT exists and that it is not learned, that is, it is not a natural language. In fact Fodor concludes that (and here we see clearly the Chomskian lineage) the LOT must be innate. (ibid., pp.79-82)

The argument obviously rests its weight on b), which is an assumption based on a belief that b) provides the best explanatory model for concept learning. The belief in turn rests on empirical evidence which fits the model (see ibid., pp.34-38). Additionally, the argument sheds light upon the form of the knowledge of conventions Fodor appeals to in the quotation above. It would seem that this knowledge amounts to the possession, by the speaker, of a theory of his or her natural language; a theory which facilitates the mapping of predicates of the natural language onto those of the LOT. It is a matter of convention which of the former will correspond to each of the latter and it is the activity of making and confirming hypotheses, of the form of 1), which establishes this correspondence.

So Nelly's understanding of Noah's locution 'It is raining' is dependent upon her possessing a theory of English which is couched in the vocabulary of the LOT. The message Noah conveys via the wave form is itself a structure belonging to the LOT and it is through the application of her theory that Nelly can translate the wave form back into the original message. The picture will be brought into full relief by exposing Fodor's assumptions, firstly, as to the nature of the vocabulary of the LOT, the elements of which will constitute the message structures traded in the course of communication; and, secondly, as to the type of translation process involved in understanding a sentence.

Let us suppose that Noah had mistakenly expressed himself by saying 'It is snowing'. Since we can say this is *not* what he intended there must be a way of differentiating between what he did and did not want to say. The way in which we, or at least he, can so differentiate is by his having a representation of what he intended to say which can be compared with what he actually did say. This representation is the message Noah had in mind and it can be compared to the wave form, or locution, via the mapping process already mentioned. By this we can adduce that the LOT possessed by language users is in fact a representational system, parts of which make up a vocabulary and, when concatenated, form message structures. This is why, in 1981, Fodor described his *Language of Thought* as propounding a Representational Theory of Mind (Fodor, 1981, p.26).

The vocabulary of the LOT consists of representations, then, and it is these which allow an agent to make comparisons between actual and intended behaviour, verbal or otherwise. The question of what would give rise to slips of the tongue such as that supposed of Noah is answered by Fodor who suggests that, 'it is plausible to hypothesize *mechanisms* of the sort whose operations would account for the respects in which the observed and the intended behaviour differ.' (ibid., p.30. My italics.) Thus, the suggestion is that behaviour is the result of mechanistic operations. To return to my trite example, the process by which the message Noah intends to communicate becomes verbal behaviour is a mechanistic one and, as in the case where what he says is not his intended locution, there can be mechanical failures (ibid.).

To summarise, the assumptions behind the account of understanding arising from Fodor's LOT hypothesis are:

1) The LOT constitutes a representational system.

2) The vocabulary of this system is innate and, hence, not a natural language.

3) The processes by which the structures of this system interact and translate into and from sentences of natural languages are mechanistic.

4) The processing occurs at an unconscious level.

The LOT, therefore, can be characterised as an innate representational system in which mechanistic operations are effected upon representational structures. It hardly need be pointed out that 'a mechanistic representational system' is often taken as an adequate definition of an electronic computer, which helps explain why the position Fodor delineated in *The Language of Thought* became the orthodoxy of Computationalism.

If the vocabulary of the LOT consists of internal representations, then the LOT can be viewed as the human equivalent of the machine code possessed by computers, and as such constitutes the formal medium within which mental transactions occur. What needs to be added is an explanation of the relation between the LOT and the natural language. Specifically, we need an account of the process of translation, of structures of the latter into those of the former, involved in understanding a proposition.

Fodor observes that general purpose digital computers usually communicate in languages different from those in which they compute, and that information passes into and out of the computational code via the operation of compiling systems which are effectively translation algorithms for the programming languages that the machine employs. As we saw in the 'Noah and Nelly' account of understanding, Fodor sees human beings as operating with two languages also, in this case English and the LOT. Similarly, according to Fodor,

> '...the mechanisms whereby human beings exchange information via natural languages....constitute "compilers" which allow the speaker/hearer to translate from formulae in the computational code to wave forms and back again.' (ibid., p.116.)

For Fodor the mind contains not only mechanisms which translate input in the form of natural language sentences into formulae of the LOT—the internal computational code—but also mechanisms which translate perceptual input (ibid., pp.117&118.). The same internal coded structure may be created by 'compilers' dealing with input from the visual, auditory, tactile , olfactory, or gustatory systems and, at least for the first three, there will be mechanisms which deal with linguistic information (heard, read, or, for the Braille reader, felt sentences) as well as purely sensory data. Nelly, for example, may have the LOT message, expressed in English as 'It is raining', engendered in her through hearing or reading Noah's warning; or through feeling, seeing, hearing, or even smelling and tasting the rain. The net result is the same; her internal code will formulate a structure which represents the state of affairs that it is raining.

Of course, since we do not have the capacity to speak and understand a natural language from birth, the 'compiler' dealing with translations from the LOT into that language and back again will have to be acquired through the process of learning the latter. It is easier to make sense of the notion of learning a 'compiler' if one acknowledges Fodor's thesis that,

> 'a compiler which associates each formula in the input language I with some formula in the computing language C can usefully be thought of as providing a semantic theory for I, taking C as the metalanguage in which the semantic properties of the sentences of I are represented.' (ibid., p.119.).

So, 'In effect, the theory of meaning for formulae in I is simply the translation function which maps them onto formulae of C' (ibid.).

My point is that by seeing the functional description of the compiler which translates English, for example, into the LOT (and back again) as a theory of English, we can explain Fodor's position thus: Learning English is a matter of learning a theory of meaning for English, and such a theory becomes, for want of a better word, 'embedded' in a compiler. This is the sense in which we might be said to learn a compiler.

At this stage we can form an outline of the complete picture Fodor paints of language learning and competence. Learning one's natural language is a matter of constructing a theory of meaning for that language which will be couched in the LOT. This theory will have been constructed through the production and selection of hypotheses regarding the truth conditions of the natural language predicates, a process which, in effect, maps these onto existing LOT predicates. Knowledge of the conventions of one's natural language is, therefore, nothing more nor less than possession of the theory of meaning for that language and the theory is realized by the compiler.

Competence in one's natural language is a result of the functioning of the mechanisms in which its theory of meaning are embedded. Sentences of the natural language are created when one uses the theory to translate from message to wave form (that is, LOT formula to spoken or written sentence), and they are understood when one employs the theory in the opposite direction. The translations are effected within the translation mechanisms, or compilers, and the functioning of these instantiates one's use of one's knowledge of the conventions of a natural language.

For a more detailed understanding of what goes on in a linguistic translation mechanism we can refer to two hypotheses presented by Fodor:

> '1. The mapping from messages to wave forms and vice versa is indirect: Wave forms are paired with messages via the computation of a number of intervening representations.
>
> 2. Among these intervening representations there are several which correspond to the structural description of sentences which generative grammars provide.' (ibid., p.109&110.)

These hypotheses, if confirmed, would entail that a translation mechanism analyses sentences of the natural language into component representations at a finite set of levels of description. Fodor draws from the work of Chomsky to posit at least the following levels; phonetic, morphophonological, surface syntactic, and deep syntactic (ibid., p.110.). He continues by suggesting that,

> 'each level of description can be identified with a certain (typically infinite) set of formulae whose elements are drawn from the vocabulary of the level and whose syntax is determined by the well-formedness rules of the level.'(ibid.)

Consequently, every legitimate sentence of the natural language is

> '...associated with a set of representations such that each formula in the set is well formed at some level of description and such that each level of description contributes at least one formula to the set. This set of formulae is the *structural description* of the sentence relative to the grammar.'(ibid.)

The translation of a sentence, then, entails its analysis into representations at various levels of description. The concatenation of representational elements in, or the syntax of, each formulae is governed by well-formedness rules at each level and the complete set of representation formulae, at all levels, for a given sentence constitutes its structural description. It follows from this that Fodor can help himself to Chomsky's idea of a generative grammar to account for a human compiler's capacity to translate an indefinite number of natural language sentences into the internal code of the LOT.

From this information concerning the translation mechanism we may affirm that the putative representational system is heterogeneous because it functions at differing levels of description. Returning to the example of understanding a spoken sentence, we find illumination regarding the nature of the representational system used in this process of cognition in the following:

> 'The perceptual recognition of an utterance involves assigning it a series of increasingly "abstract" representations (one for each level of linguistic description acknowledged by the grammar of the language)...'

Producing an utterance, on the other hand,

> '...involves representing the intended [verbal] behaviour as satisfying the corresponding series of decreasingly abstract representations, the last member of which can be read directly as a phonetic matrix.'(ibid., p.158)

Recalling the Chomskian levels of description just mentioned, we can deduce that 'concrete' representations are to be found at the acoustic/phonetic level of sentence analysis, while 'abstract' ones are employed at the deep syntactic level (see ibid., pp.160&161 for support of this deduction).[16] So, for a person to fully understand a sentence he or she will need to produce a full structural description of that sentence which will comprise various formulae consisting of representations at differing levels of abstraction.

So, the LOT processes discursive stimuli and responses by employing a variety of levels of representations. However, Fodor does not believe that the representations comprising the LOT are solely discursive. He thinks that there is evidence that besides the existence of discursive representational mechanisms in people there is a capacity for imagistic representation which is central to a variety of cognitive functions. That said, for Fodor the use of images in cognitive processes is parasitic upon discursive representation in the sense that images are constructed to accord with descriptions, 'That is', he writes, 'we have access to a computational system which takes a description as input and gives, as output, an image of something that satisfies the description'.[17] Imagistic representations, thus, would be employed in cognitive processes when the characteristics of the cognitive task demanded it (ibid., p.192).

[16] Presumably, at the surface level, there will be 'concrete' representations of pitch and tone as well as of phonemes present in the wave form. The 'abstract' representations, perhaps, would be akin to the phrase markers which appear at the deep structure level of analysis in Chomsky's Theory of Transformational Grammar, the level at which semantic interpretation is fully effected.

[17] Fodor is aware that an explanation of cognition solely in terms of the manipulation of images is untenable. The objections to such an explanation are too strong.

One objection is that the references of images are ambiguous. This is evident if, for example, the representational system is to

Fodor's LOT, then, though capable of employing imagistic representations, is primarily discursive. That is, it is structured as a language and, therefore, the representations it employs are subordinated by sets of linguistic rules, or grammars. Of course, these rules themselves have representation within the system, along with the images, but the latter have no utility within the system until they are assigned a functional role by the former.

In the light of Fodor's insistence on the linguistic basis of the functional organisation of the mind, we can venture beyond my original exemplar of understanding a sentence to look, briefly, at other aspects of Fodor's theory of mentation. Returning to Nelly, once again, it is reasonable to say that on understanding Noah's message she will believe it is raining and, supposing it is raining and her belief is justified, she will also know it is raining (if we ignore the blight of Gettier counter-examples). Noah's locution, then, has lead to the instigation, in Nelly, of a cognitive process (her understanding) and two cognitive states (her believing and her knowing it is raining). It might also lead her to wish it was not raining, hope it will stop, worry that the animals will get wet, and much more besides. For Fodor, all of these psychological states are explicable in terms of the LOT hypothesis because they can be individuated in such a way that their descriptions will include a linguistic structure, with propositional content, which can be represented in the internal code. In the examples just mentioned the propositions could be 'It is not raining', 'It will stop raining', and 'The animals will get wet', respectively. As natural language sentences these are all amenable to an analysis

replace the sentence 'John is fat' with an image of the generously proportioned John. An image of John may be required to represent the sentence 'John is tall' or 'John is hirsute', perhaps even 'John is a Mormon', and since the same image would suffice as a representation of each sentence the system will be incapable of differentiating between them. Clearly a system of representations which can not distinguish between sentences with wholly different meanings can not be cognitive.

Secondly, mental images are indeterminate. To adapt Dennett's example (Dennett, 1969, pp.136&137), if the production of the utterance 'Look out, there's a Tiger!' entails the representation of the beast as an image it ought to be possible to count how many stripes it has in the image. It ought to be possible because one's use of the word 'tiger' is referential and, since the tiger referred to has a determinate number of stripes, the image, if it is to refer, must also be determinate. However, few people could honestly claim to have such determinate images which invites the conclusion that an imagistic representational system would not be capable of supporting cognitive functions.

Fodor gives mental images a subordinate role in the LOT and thereby rescues them from explanatory oblivion. He suggests that when derived from descriptions the images employed in the LOT would be disambiguated and determinate thus affording them reference (thus securing them truth-values).

in the LOT which will terminate in a structure representing the state of affairs described in each.

The exegesis is, so far, incomplete. This becomes apparent if we reconsider Nelly's states of belief and knowledge. Nelly can believe it is raining when it is not and therefore can be said not to know it. Clearly, belief and knowledge are different states. However, at this stage the explanation does not differentiate between them because they can both be described as representing the same proposition. What does distinguish them is the relation Nelly has to the proposition or, put another way, her attitude toward it, hence the phrase 'propositional attitude' which is used to categorise this type of psychological state. 'So', Fodor explains, 'having a propositional attitude is being in some relation to an internal representation'. In answer to the question 'What sort of relation?' he writes, 'In particular, having a propositional attitude is being in some *computational* relation to an internal representation' (ibid., p.198). Fodor expands upon this by saying,

> 'Mental states are relations between organisms and internal representations, and causally interrelated mental states succeed one another according to computational principles which apply formally *to the representations.*'(ibid.).

Thus, Nelly's understanding of Noah's sentence would cause her to believe it is raining when the message representation undergoes a computational process the principle of which is, presumably, to be discovered by cognitive theorists.

For now, we can say that a propositional attitude is a representation computationally, and causally, related to other representations. The representation's causal role will partly determine the propositional attitude with which it is associated, but the representation must be computationally derived from the LOT as a formula within it. This last stipulation is important as a constraint upon any cognitive theory since it highlights the need for such theories to ensure that the content of a representation is synchronised with its causal efficacy as both determine the assignment of propositional attitudes to an organism (see ibid., p.199). To use Fodor's example, a theory of cognition might assign the English sentence 'There aren't any aardvarks any more' to a physiological state nomologically necessary and sufficient for (or contingently identical to) the belief that it is raining. This assignment would be a consequence of the computational model the functional organisation of which makes it the case that the physiological state in question must represent the aforementioned sentence. Of course, the assignment is unacceptable because, as Fodor puts it, '...the *causal* consequences of believing that it will rain can't be paired in any coherent way with the *logical* consequences of "There aren't any aardvarks any more".'(ibid.).

It would be odd if a theory posited a law-like causal connection between Nelly's behaviour of reaching for an umbrella and the physiological state representing the sentence 'There aren't any aardvarks any more'. Odd because it would contravene the obvious requirement that the semantic content of a mental state be logically connected to the behaviour it brings about. Fodor's account of the mind's functional organisation is intended as a preliminary to a computational model which would guarantee such logical connections were maintained in its processing.

4 THE ONTOLOGY

The chapter began by sketching an historical thesis concerning recent philosophies of mind. The thesis held that there is a linear progression through the positions of the Cartesian Dualist, the Behaviourist, the Identity Theorist, to the position of the Computationalist. The movement is seen as progressive because in it the problem as to the ontological status of the mind receives an increasingly more sophisticated treatment.

Putnam's conclusion in the sixties was that the mind-body problem is a philosophical aberration because we can describe the functional organisation of the mind—what it does and how it does it—without need of an ontological stance. As I have noted, he abjures any notion of mental type to physical type reduction.

Fodor also rejects reductionism. He does so on the grounds that he finds it implausible to expect psychological laws to reduce to physical ones. The implausibility is due to the unlikelihood that the kind predicates, projected by psychological laws, can be paired, by bridging laws, with the kind predicates of physics.[18] Put another way, Fodor believes that the two types of kind predicate are not coextensive (see ibid., p.9 ff.).

To illustrate this point Fodor takes an example from another of, what he calls, the special sciences (as opposed to the basic science of physics), namely, economics. A law of economics which includes as a kind predicate 'is a monetary exchange' would, if it is to be reduced to a law of physics, require bridging laws expressing contingent identities of the form 'physical event P is a monetary exchange'. However, such identification could not be made between a physical kind and this economic kind because the extension of the predicate 'is a monetary exchange' can include many disparately constituted physical events. Compare putting coins in a vending machine with handing over notes to a shopkeeper, or sending a cheque by post, or giving credit card details over the phone, for example. It is hard to conceive of any physical kind predicate which would be applicable to all these events and any others that can be described as a monetary exchange. As Fodor asks 'What are the chances that a disjunction of physical predicates which covers all these events...expresses a physical kind?' (ibid., p.15).

[18] By 'kind predicates' Fodor presumably means the descriptive expressions which pick out natural kinds. A natural kind is usually understood as an object, or event, type which can be individuated by a characteristic property, or group of properties. It is usually assumed that these properties can be explained by the kind's ultimate physical constitution since this determines the object's, or event's, properties and principles of functioning. For example, as a natural kind water has the ultimate constitution of being H_2O which accounts for its properties of fluidity (at normal atmospheric temperatures and pressures), transparency, and lack of odour and colour.

It is the requirement that every natural kind have an ultimate constituency, or essence, that makes the concept of a 'natural kind' a feature of essentialist philosophies (as expounded by Putnam, Wiggins, and Kripke).

Though Fodor's dismissal of reductionism entails a denial of *Type* physicalism (see footnote 5) he allows *Token* physicalism to be a valid ontological stance. A psychological kind predicate can describe events that are physical even though these events are unlikely to possess properties with the homogeneity required for them to characterise a physical kind predicate. Understanding the sentence 'It is raining' may be a physical process, but it need not follow that 'has understood the sentence' is a predicate coextensive with a physical kind predicate like 'has undergone P', where P is a physical process. A token of a psychological kind may be identified as a token of a physical kind without an entailment that the kind predicates be coextensive.

Thus, Fodor validates a form of physicalism, albeit a weaker form than reductionism. Indeed, he seems compelled to adopt a physicalistic stance on the question of the mind's ontological status since he thinks that a mind is a computer housed in an organism where its states are causally implicated in the production of behaviour and, as he says,

> 'When we think of an organism as a computer, we attempt to assign formulae in the vocabulary of a psychological theory to physical states of the organism (e.g. to states of its nervous system).'(ibid., p.73).

After all, it is only when the mind's computational states are given a physical description that the supposition that they *cause* behaviour becomes coherent. The notion of causation Fodor is applying, then, would seem to be that of physical, nomic, causation.

Further support of the view that Fodor adopted a physicalistic position, in *The Language of Thought*, can be found in the following:

> 'The idea is that, in the case of organisms as in the case of real computers, if we get the right way of assigning formulae to the [physical] states it will be feasible to interpret the sequence of events that *causes* the output as a computational *derivation* of the output. In short, the organic events which we accept as implicated in the etiology of behaviour will turn out to have two theoretically relevant descriptions if things turn out right: a physical description by virtue of which they fall under causal laws and a

> psychological description by virtue of which they constitute steps in the computation from the stimulus to the response.' (ibid., pp.73&74).[19]

In order to fully appreciate Fodor's conception of the enterprise of cognitive science it is helpful to consider his avowed assumption

> '...that psychologists are typically in the business of supplying theories about the events that causally mediate the production of behaviour and that cognitive psychologists are typically in the business of supplying theories about the events that causally mediate the production of intelligent behaviour.'(ibid., p.9).

Fodor's LOT hypothesis, then, is, in part, an attempt to outline how psychological processes, such as reasoning or thinking, can be given a role which explicates how they can cause behaviour. The role he gives them is computational and as causes they are to be seen as physical. To bring us up to date, in his recent book *Concepts: Where Cognitive Science Went Wrong* Fodor entwines the representationalist, computationalist, and physicalist threads saying;

> 'In a nutshell: token mental representations are symbols. Tokens of symbols are physical objects with semantic properties. To a first approximation, computations are those causal relations among symbols which reliably respect semantic properties of the relata' (Fodor, 1998, p.10)

Since Fodor presents LOT as a model of the basic architecture of the human mind he is assuming more than a simple functional isomorphism between the model and the mind. Whereas Putnam, in comparing the mind to a Turing machine, might be taken to be suggesting little more than that humans cogitate by passing through computational, or functional, states,

[19] This passage occurs as part of an explanation why the LOT hypothesis should not fall foul of objections stemming from Wittgenstein's Private Language Argument (see Wittgenstein, 1953, §258 ff.).The argument raises doubt as to whether terms of a private language can be used coherently since there is no independent criterion for judging correct from incorrect usage (independent, that is, of the user's beliefs on this matter). Fodor's contention is that coherence is guaranteed a language if there is a correspondence between intentions, and other propositional attitudes, and the linguistic forms used to express them. In the case of a public language this correspondence is apparently secured by conventions for the use of terms but, in the case of the LOT, the connection will be effected by assigning formulae to causally implicated physical states in such a way that the states' causal relations will guarantee the logical relations amongst the formulæ (Fodor, 1975, pp.70-75).

So, Fodor's defence against Private Language Argument objections relies on the assumption that the coherence of a language (for example, consistency in the use of its terms) could result from its being generated mechanistically, in this sense, according principles of cause and effect.

Fodor's thesis goes a lot further. In it he offers, admittedly rough, sketches of the sort of states (and the structures in which they inhere) we might one day discover in the mind. Beyond this he intimates that the states and structures will be shown to be physically instantiated. Computers endowed with a LOT and its attendant functional capacities would not only be able to reproduce intelligent behaviour, but would generate it in the way we do. A thesis which suggests computers may mirror human linguistic performance *and* competence must rest in the strong AI camp and deserve the epithet ‘Classical Computationalism’.

CHAPTER TWO

ELIMINATIVISM

In the last chapter I drew attention to a bifurcation in approaches to the philosophy of mind which emerged from the Central State Materialism of the 1950's. One branch of this fork was to grow, via Functionalism, into the Classical Computationalism of Fodor whilst the other developed into Eliminative Materialism. What is interesting about these two branches is that despite sharing the same physicalist roots (for both hold that cognitive activity occurs in a physical medium) their proposed explanatory schemes are mutually exclusive. I shall say more about this schism in due course but will begin by providing an outline of the Eliminative Materialist thesis.

The Type Identity Thesis holds that certain types of mental state, process, or event, are identical to types of neural state, process, or event, which will be individuated by neuroscientists at some time in the future. Another way of stating this thesis is to say that for certain mental predicates there are corresponding ones to be found in neuroscientific vocabulary with which they, one day, will be shown to be coextensive. Stated in this way it is possible to see the thesis as committed to the view that mental predicates denote natural kinds occurring in the world which, it so happens, are also denoted by neural predicates. It is an ontological commitment not only to a monistic stance with regard to substance but also, less obviously, to the existence of mental phenomena. When one says of another 'He is in pain' or 'He saw my bacon sandwich' one is using the words 'pain' and 'saw' to describe actual states of pain and seeing. Eliminativism takes this second commitment to be misguided.

1 PAUL CHURCHLAND

a) Non-Reductive Materialism

Psychological terms such as 'pain', 'seeing', 'belief', and 'desire' are commonly taken, among philosophers of mind, to be items in the vocabulary of what they call 'folk psychology' and this, in turn, is taken to constitute a common-sense framework shaping our understanding of human behaviour. If, as Eliminativists suggest, folk psychology is to be viewed as a theory of mind, then the Identity Theorist can be said to be proposing an intertheoretic reduction of folk psychology to neuroscience. Such a reduction requires that the predicates of folk psychology match up, one-to-one, with the predicates of neuroscience. It is this consequence that Eliminativists will not countenance. Paul Churchland puts the case as follows,

> 'As the eliminative materialists see it, the one-to-one match-ups will not be found, and our common-sense psychological framework will not enjoy an intertheoretic reduction, *because our common-sense psychological framework is a false and radically misleading conception of the causes of human behavior and the nature of cognitive activity*....Accordingly, we must expect that the older framework will simply be eliminated, rather than be reduced, by a matured neuroscience.'(Churchland, P.M.,1988, p.43.).

Fodor's Classical Computationalism and Churchland's Eliminative Materialism may be taken, at first sight, to concur in their rejection of reductionism. After all, both would agree that physical kind predicates will not occur in identity statements with psychological kind predicates, whilst adhering to a physicalistic ontology. However, the concurrence is superficial for Fodor maintains that this psychological-physical kind predicate incommensurability requires us to bestow upon psychology an autonomic status (because the incommensurability precludes the subordination of psychological by physical laws) in contrast with Churchland who believes it should lead us to altogether abandon psychology (at least in any recognisable form).

Since Churchland believes that the extension of kind predicates is fixed by the theoretical framework in which they originate (see ibid., pp.56-59), we can see why, for him, the rejection of the common-sense psychological framework entails a refusal to accept that psychological predicates have any extension at all. If, for example, the extension of the predicate ‘is in pain’ is specified only by virtue of its occurrence in nomic statements of the folk psychological theory (such as ‘persons in pain tend to want to relieve that pain’) then it is reasonable to conclude that should the theory, with its constituent nomic statements, be invalidated then the question of what ‘pain’ refers to will be empty. Here Churchland's Eliminativism stands in contrast to both the Identity Thesis and Fodor's Computationalism in its denial of the reality of psychological phenomena.

Eliminative Materialists are eager to assure us that the removal of entities from our everyday, as well as scientific, ontology is a common enough practice. Rorty draws attention to a number of examples of such ontological elimination. He cites the replacement of ‘a quantity of caloric fluid’ by ‘mean kinetic energy of molecules’ in explanations of a body's temperature as one example, the explaining away of witches as psychotic women as another, and, as a plausible hypothetical case, the dismissal of accounts of illness in terms of demonic possession in favour of theories of pathology which contributed to the elimination of demons from our world view. In each case a newer more powerful mode of explanation has replaced an old theoretical framework and in replacing it has rendered some, or all, of its substantival expressions obsolete. Rorty's contention was that the same fate could, at least in principle, befall our use of sensation-discourse (Rorty, 1965, pp.19-21).

Paul Churchland also cites the first and second of these examples of elimination and adds one or two of his own. His historical precedents are phlogiston, a spirit-like substance supposedly found escaping from wood when it burns and metal when it rusts, and the starry sphere of the heavens believed to exist by early astronomers. 'Phlogiston' was deprived of its role as a designating expression by pneumatic chemistry whilst 'the starry sphere' was similarly deprived by Kepler's theory of planetary motion (Churchland, P.M., 1988, p.44). That the terms employed by folk psychology might suffer a comparable fate gains credence from Churchland's observation that in primitive cultures the behaviour of natural phenomena was explained using such terms. For example, the wind could *know anger*, the moon *jealousy*, the river *generosity*, and the sea *fury* (Churchland, P.M., 1981, p.211). The subsequent ability to explain the activity of natural phenomena as in terms of physical causation robbed them of agency and made attributing psychological states to them seem, at best, fanciful. The Eliminativist contends that, in principle, a similar explanation of human activity could have the same consequence.

However, Eliminative Materialists do not merely contend that, *in principle*, the vocabulary of folk psychology *could* become obsolete, for they are inclined to make the stronger claim that, *in practice*, this vocabulary *will* become obsolete. Admittedly they usually qualify this claim by saying that the proposed elimination is subject to the empirical inquiries of neuroscientists, but they offer several reasons why we should expect these inquiries to find against the views implicit in the folk psychological outlook. There is, however, an important assumption which requires examination prior to attending to those reasons.

b) The Folk Theory

Most of Churchland's argumentation for the future elimination of folk psychology rests on the premiss that it constitutes a theory. In *Scientific Realism and the Plasticity of Mind* (Churchland, P.M., 1979) he gives support to this premiss by arguing that we need to appreciate that folk psychology is a theory if we are to understand how it contends with the problem of other minds. This requires a little explanation.

The problem is epistemological in nature since it arises from the perceived need to explain our epistemic access to the psychological states of others. Clearly this access is not via introspection or any other direct route since few of us would claim to be 'mind readers' and, question begging notwithstanding, it would be false to say we know the thoughts of others by observing their neural activity. So how does one know what others are thinking, feeling, wishing, deciding, and so on? One traditional answer is that one knows by analogy with one's own case. If another behaves in a way analogous to the way one behaves when experiencing pain, for example, then it seems justifiable to infer that the other is also experiencing pain.

It is noteworthy that the argument from analogy was promulgated by John Stuart Mill (Mill, 1889) who was also responsible for a proposed resolution of the problem of induction (which is, again, epistemological in nature since it arises from the need for a justification of the practice of generalising knowledge of particular cases). However, as an example of inductive reasoning itself, the argument from analogy leaves much to be desired for it amounts to a generalisation from only one case, *viz.*, one's own.

Churchland does not cite this unfounded inductive leap as the argument's weakness saying instead that, 'the problem with the argument from analogy is that it accedes in the representation of one's knowledge of other minds as being essentially parasitic on one's knowledge of one's own mind'(Churchland, P.M., 1979, pp.90&91). Although he does not expand upon this problematic accession I think Churchland might have been gesturing in the direction of the Wittgensteinian insight that one can not be said to *know*, from one's own case, what feelings, thinking, and other mental occurrences are. The insight, which I will refrain from expounding here, is put to use by Norman Malcolm who demonstrates how, in grasping it, one not only sees the folly of the argument from analogy but also that the reasoning giving rise to the problem of other minds is erroneous (Malcolm, 1958). Churchland also believes that the problem is chimerical but, unlike Malcolm, still perceives a need for an account of how it is we are able,

> 'to explain, predict, and understand the behaviour of certain animated particulars in terms of the wants, beliefs, pains, cogitations, and other psychological states and sequences to which they are presumed subject...'(ibid., p.91).

So, how do we justify the making of judgements about other minds? The only sensible way to answer this question, for Churchland, is to say that our judgements about the psychological states of others are the outcome of the application of a *theory* which allows us to infer their existence from behaviour and circumstances. Thus, we should understand our psychological generalisations as forming a theory 'whose credibility is a direct function of how well it allows us to explain and predict the continuing behaviour of individual human beings'(ibid., p.91).

It is a pre-requisite of this understanding of the way in which we make psychological ascriptions that there is in each of us a,

> 'tacit understanding of a framework of abstract laws or principles concerning the dynamic relations holding between causal circumstances, psychological states, and overt behaviour'(ibid., p.94).

Our ability to predict, explain, and understand, the behaviour of other human beings and, to a lesser extent, animals is, for Churchland, a consequence of our having learnt a structure of generalisations, laws, or principles, which together constitute a theoretical framework.

Churchland furnishes us with examples of the sorts of generalisation, principle, or law, he believes constitute the person theory of humans or P-theory, as he abbreviates it (ibid., pp.92&93). Here is a selection:

1) Persons tend to feel pain at points of recent bodily injury.

2) Persons in pain tend to want to relieve that pain.

3) Persons believing that P, where P elementarily entails Q, tend to believe that Q.

4) Persons subject to a sudden sharp pain tend to wince and/or cry out.

5) Persons who are angry tend to frown.

These generalisations are intended only to express gross regularities. In the case of 1) the regularity is between an external circumstance type and an inner effect type; in the cases of 2) and 3) it is intra-mental regularities that are being expressed; in the cases of 4) and 5) inner event types are being cited as the regular causes of certain forms of overt behaviour.

So, in Churchland's view, it is our tacit understanding of generalisations, such as those above, which constitutes our knowledge of the P-theory, and it is this knowledge which allows us to explain and predict the behaviour of other humans. One's attribution of minds to others is not based upon the assumption that since they behave in an analogous way to oneself others must have similar inner, mental, causes of their behaviour. Instead, the attribution is a consequence of one's having learnt a theoretical framework with which to hypothesise as to the causes and effects of the behaviour of others. Although Churchland does not say so himself, it seems fair to extrapolate from what he does say that the knowledge that others have minds is as secure only as the theoretical framework by which it is implied. Before giving Churchland's reasons for doubting this security it would be as well to consider an objection, which he raises and dismisses, regarding his insistence that our everyday use of psychological predicates presupposes our possession of a theoretical framework from which their significance is derived.

As we have seen, according to Churchland, propositions about the P-states of others, that is, third person ascriptions of psychological states, arise from a common sense theory which he was to call, in later writings, 'folk psychology'. An objection to the interpretation of psychological concepts as theoretical emerges when we turn our attention to first person ascriptions. Such ascriptions, it might be argued, are not made on the basis of inferences from general speculative assumptions like 1) to 5) above. For example, when one says 'I am in pain' one's assertion is not inferred from a general proposition such as 1) above. Rather, or so the argument runs, the assertion is based on introspective observation which is direct and non-inferential. If so, then the concept of pain is being applied in a non-theoretical context. What goes for pain would also go for any other psychological concept with a first person application—a consequence which undermines the claim that these are, in fact, theoretical concepts because it is difficult to see how the same concept can be both theoretical and non-theoretical.

Churchland's response to the objection is to assure us that though we can ascribe P-states to ourselves non-inferentially, we must do so using concepts semantically embedded in the P-theoretical framework. This is because the non-inferential judgements we make, based on the experiential qualities of states revealed by introspective observation, are not generally necessitated by any intrinsic experiential feature of those states. The concepts we apply in these judgements acquire their sense, not from the intrinsic qualities of the experiences they are used to describe, but from their role in the theoretical framework we use to interpret those qualities. Thus, when, on the basis of introspective observation, one judges 'I am in pain' the sense of the word 'pain' is derived from this framework (see Churchland, P.M., 1979, sections 2. and 3., and 1988, p.56).

If we accept Churchland's account of the semantics of observational concepts then the objection is dissipated,

> 'For there is nothing inconsistent in the idea that one should be able to make reliable non-inferential applications of a concept whose semantic identity is fixed by a theory. All one needs to do is continue a reliable habit of conceptual response to situations where the concept at issue truly applies.' (Churchland, P.M., 1979, p.96).

For Churchland the habit is initiated by the infant's acquisition of the psychological vocabulary which amounts to its initiation into the explanatory and predictive uses of the P-theory. He urges that,

> 'the infant has as great a need to generate or acquire a conceptual framework with which to comprehend his own states and internal activities as he has to generate or acquire a conceptual framework with which to comprehend the states and activities of the world at large.'(ibid., p.99).

The ability to make introspective judgements is, therefore, a further consequence of acquiring the P-theory and these judgements, being theoretical in nature, are corrigible, dubitable, and fallible.

So, having seen how Churchland validates his claim that the ordinary psychological predicates we use are theoretical, and how he defends the claim against the argument that this can not apply to first person ascriptions of those predicates we are now in a position to examine his arguments why we should expect the P-theory, or folk psychology, to be eliminated by a matured neuroscience.

c) Reasons to Reject the Theory

What reasons, then, does Churchland give for thinking that folk psychology, as a theory, will be eliminated? By way of answer to this question I shall briefly outline six lines of argument which occur, often recurrently, in Churchland's writings.

1) Although he concedes that folk psychology 'does enjoy a substantial amount of explanatory and predictive success' Churchland points to what he identifies as severe limitations in its explanatory power (Churchland, P.M., 1981, p.210). Apparently, for examples of this deficiency we need only consider its inability to explain the nature and dynamics of mental illness, the faculty of creative imagination, the grounds of intelligence differences between individuals, the nature and psychological functions of sleep, the ability to catch a baseball whilst running, the internal construction of a 3-D image from subtle differences in the 2-D array of stimulations on each of our retinas, the rich variety of perceptual illusions (visual and non-visual), or the miracle of memory with its lightning capacity for relevant retrieval (Churchland, P.M., 1981, p.210, and 1988, pp.45&46). There is also a conspicuous lacuna in folk psychological explanation regarding the process of learning, especially where this involves large-scale conceptual change and as it occurs in its pre-linguistic or entirely non-linguistic form (as in infants and animals) (Churchland, P.M., 1981, pp.210&211). As we shall see in the sixth line of argument below, the existence of this lacuna is an indication of a flaw in folk psychology revealing more than explanatory weakness.

2) Borrowing Imre Lakatos's terms, Churchland accuses folk psychology, or FP, of being a 'stagnant and degenerating research program' (ibid., p.211). That the theory is degenerate is illustrated by the retreat in the scope of application of the psychological, or intentional, idiom from multifarious natural phenomena to the much narrower domain of the higher

animals (for, as was noted earlier, we no longer speak of the anger of the wind or the generosity of the river). Regarding the latter Churchland writes,

> 'Even in this preferred domain, however, both the content and the success of FP have not advanced sensibly in two or three thousand years. The FP of the Greeks is essentially the FP we use today, and we are negligibly better at explaining human behavior in its terms than was Sophocles.'(ibid.).

Given the backlog of anomalies and mysteries in its explanatory domain such stagnation in a theory is inexcusable.

3) Folk psychology, Churchland points out, fails to cohere 'with the rest of our developing world picture'. The world picture of which he speaks is that constructed by the physical sciences such as particle physics, atomic and molecular theory, organic chemistry, evolutionary theory, biology, physiology, and neuroscience. He complains that folk psychology 'is no part of this growing synthesis. Its intentional categories stand magnificently alone, without visible prospect of reduction to that larger corpus.'(ibid., pp.211&212). It is the unlikelihood of a future reduction into neuroscientific categories that gives folk psychology a comparable aspect to alchemy, and vitalism, both of which were eliminated by subsequent sciences.

4) The fact that so many 'folk' theories have been eliminated in the past ought, Churchland maintains, to point us, via a little inductive reasoning, to the conclusion that our sole remaining folk theory, the P-theory, will be eliminated. He cites early theories of motion, the structure and activity of the heavens, and the nature of fire, and of life, as instances of eliminated folk theories which should lead us to expect that folk psychology will face a similar fate (Churchland, P.M., 1988, p.46).[1] This is particularly likely in the case

[1] Churchland does not make it entirely clear what we are to count as a 'folk' theory. Should we, for example, consider Newtonian mechanics or Boyle's gas laws as constituting such a theory? Both are given, by Churchland, as instances of 'humble theories', parallel with folk psychology, which flounder when applied to 'unexplored dimensions of their old domain' (these being velocities close to the velocity of light in the case of Newton's theory, and high temperatures and pressures in the case of Boyle's) (see Churchland, P.M., 1988, p.46).

If we are to count these as folk theories then must we not attribute that status to current theories also? Theories produced by of folk psychology, he contends, because of the complexity of the phenomena of conscious intelligence which it attempts to systematically conceptualise. Indeed, 'So far as accurate understanding is concerned, it would be a *miracle* if we had got *that* one right the very first time, when we fell down so badly on all the others' (ibid.).

5) Churchland switches from a posteriori to a priori reasoning to further encourage us to anticipate the vindication of his Eliminativism. Assuming that a materialistic account of mental phenomena will be forthcoming, he concludes that there is a greater a priori probability of Eliminative Materialism providing that account. The two principle rivals for explanatory success are the Identity Theory and Functionalism and both, according to Churchland, wager that the concepts of folk psychology will match-up with concepts from a matured neuroscience.[2] However,

> 'there are vastly many more ways of being an explanatorily successful neuroscience while *not* mirroring the structure of folk psychology, than there are ways of being an explanatorily successful neuroscience while also *mirroring* the very specific structure of folk psychology.'(ibid., p.47).

Consequently, 'the a priori probability of eliminative materialism is not lower, but substantially *higher* than either of its competitors.'(ibid.).

6) If we wish to provide a theory of the cognitive development of humans from infancy to adulthood, and we also wish to integrate propositional attitude ascriptions of folk psychology within this theory, then the cognitive state and process attributions indicative of the development will involve the attribution of propositional attitudes to infant and adult

contemporary physicists and, of course, neuroscientists would have the same status as the P-theory and, given Churchland's inductive argument, should also be expected to be eliminated. It would seem that without careful, and somewhat arbitrary, definition of the sense of the word 'folk' in this context Churchland would be throwing out the baby with the bathwater.

[2] Churchland concedes that the Functionalist, in denying the existence of universal type/type identities, would expect the relevant match-ups to be merely species-specific, or even person-specific (ibid., p.47). However, as we saw in the last chapter, neither Fodor nor the early Putnam presume psychological type/physical type identities, whether species or person-specific. For Fodor the unlikelihood that psychological kind predicates will reduce to physical kind predicates persuaded him that while token physicalism is a tenable position, the Type Identity Theory is not. For Putnam the ontological question, of what the mind is, was irrelevant to his Functionalist thesis. Indeed, part of the appeal of Functionalism was that it was not necessary to match-up psychological with neuroscientific concepts but only psychological with ontologically neutral functional.

Churchland's argument can be aimed at the Type Identity Theory but not Functionalism. Any fortification it gains for Eliminative alike. Churchland would consider such a theory to be an instance of the Ideal Sentential Automaton (ISA) approach which he characterises as follows;

> 'On the ISA approach we assume that, at least for the purposes of normative theory, the current state of an epistemic engine is relevantly and adequately represented by a set of *sentences* or *propositions*.... We assume also that the epistemic system is subject to inputs, *also representable by sentences*, which can be of two distinct kinds: (i) fresh observations, and (ii) new hypotheses.' (Churchland, P.M., 1979, p.125).

Such an approach is, in Churchland's opinion, untenable because any theory it might produce will fall foul of the argument I shall outline.

We begin with two premisses:

a) Rational intellectual development in an infant cannot be properly represented by a sequence of suitably related sentences. Put simply, a pre-linguistic infant cannot be attributed with propositional attitudes.

b) The rational intellectual development of infants is continuous with later, linguistic, stages of development. Continuity pervades the dimensions of behavioural, structural (that is, physiological structures), and functional (in particular the physiological functions underlying the processing of information) development (see ibid., pp.133-136).

From these we draw the conclusion:

c) 'Sentential parameters cannot be among the primitive parameters comprehended by a truly adequate theory of rational intellectual development'. That is, we cannot expect to find, in the developing brain, structures corresponding to propositional attitudes, which suggests that their theoretical relevance is superficial, or at best derivative, even in the case of language using adults.(ibid., p.128).

Materialism it will also gain for Functionalism.

So, if Churchland's argumentation is sound, folk psychological theory does not fit the facts about intellectual development.

The foregoing arguments are intended to undermine a position which maintains that any complete characterisation of the human behaviour must account for the phenomena identified by the vocabulary of folk psychology. The position is in direct opposition to Eliminative Materialism because it takes an ontological stance from which propositional attitudes (and, of course, sensations and perceptions) are seen as existing in the world alongside tables, chairs, force, mass, and acceleration. For this reason the position has come to be known as 'Realism' (though it is to be distinguished from the more broadly defined 'realism' which holds, very roughly, that there is a mind-independent reality, for Eliminativists usually are realists in this sense) and it is occupied by such antagonistic bed-fellows as the Dualist, the Type Identity Theorist, and the Classical Computationalist. Of course, the antagonism is more notable than the agreement, particularly when we consider how each regards the possibility of the reduction of folk psychological to neurological theory.

A Dualist would deny the possibility of reduction simply because the phenomena described by one theory are of a different ontological category from those of the other. The Identity Theorist, in contrast, would argue that for each type of event or state individuated by the theory to be reduced there is a type of event or state with which it may be identified in the reducing theory. As we saw at the end of the last chapter, Fodor, the Classical Computationalist, maintains that folk psychology can not be reduced to neurophysiology although the phenomena identified by the former are not of a different ontological category from the latter (for Fodor seems inclined, and, in my view, compelled, to anticipate the vindication of a Token Identity theory at some future date). However, if any or all of

Churchland's arguments do herald the demise of folk psychology, then all three occupants of the Realist position are discredited simultaneously.

Dualists and even Type Identity Theorists have become increasingly more difficult to find. The problem of explaining the interaction between physical and non-physical events has always been the bane of Dualism and despite brave, and ingenious, attempts to overcome it (see, for example, the 'Microsite Hypothesis' advanced in Eccles, 1987) the problem seems insurmountable. In a time when a favoured gauge of philosophical respectability is the extent to which a thesis follows the contours mapped by the physical sciences, the postulation of non-physical mental events may appear somewhat quixotic.

Cognitive scientists may have felt inclined to commend the Type Identity Theorists for re-centralising at least some mental phenomena after the Behaviourists had driven them out to the peripheries of stimuli and responses, and for insisting that the ethereal properties of mental events are, in fact, properties of physical events. However, the same cognitivists are likely to consider the Identity Theory somewhat naïve in its proposed straightforward reduction of the mental to the physical (we need only recall Putnam's point that the same, functionally described, mental event type could be instantiated in a creature with a very different internal constitution from our own, making a type-type identification impossible). More relevantly, the Identity Theorists, as was noted in the previous chapter, seemed reluctant to bring cognitive concepts within the bounds of their central state materialism.

So when an Eliminativist like Paul Churchland attempts to undermine Realism his arguments threaten to bring down three potent approaches to explaining behaviour. Given that our interest in this study is the assessment of attempts at accounting for cognition, and of the three approaches it is that of the Classical Computationalist which specifically aims to provide such an account, I shall proceed by examining further the contrariety of Eliminativism to this approach.

2 PATRICIA CHURCHLAND

a) The Folk Theory

Patricia Smith Churchland is also an occupant of the Eliminative stance which she propounds in her book *Neurophilosophy* (Churchland, P.S., 1986). At first sight this work may appear to be at variance with the work of Paul Churchland because a good deal of its philosophical content is devoted to vindicating the thesis that psychology can and will be reduced to neuroscientific theory (see Part II of the book). Before examining the form of reduction Patricia Churchland envisages for psychology we need to acknowledge an important premiss upon which her reasoning is based.

The premiss to be acknowledged is the, by now, familiar one that folk psychology constitutes a theory. The importance of the premiss for Churchland is that, clearly, without it there can be no argument for inter-theoretical reduction from folk psychology to neuroscience. The premiss is made explicit in the following characterisation:

> 'Folk psychology is commonsense psychology - the psychological lore in virtue of which we explain behavior as the outcome of beliefs, desires, perceptions, expectations, goals, sensations, and so forth. It is a theory whose generalizations connect mental states to other mental states, to perceptions, and to actions. These homey generalizations are what provide the characterization of the mental states and processes referred to; they are what delimit the "facts" of mental life and define the explananda.' (ibid., p.299).

So, folk psychology offers explanations of behaviour based on generalisations connecting mental states to perceptions and actions. Churchland gives the following as an example of such a generalisation:

1) 'Whenever a person wants to bring about a state *s*, *and* he believes that his doing *p* is a way to bring about *s*, *and* he believes that he can do *p*, *and* he can do *p*, then, barring conflicting desires or preferred strategies, he will do *p*.' (ibid., p.300).

She suggests that this might be an unspoken, or 'understood', generalisation but one which is, nevertheless, implicit in our explanation of some behaviour—my pouring a glass of water to quench my thirst, for example. Whether unspoken or otherwise, the generalisation purports to describe a regularity in nature and is, therefore, properly seen as theoretical and on a par with a generalisation asserting a regularity between a piece of copper being heated and its expanding.

So far the Churchlands' accounts of folk psychology are in close correspondence. The correspondence continues with respect to their defences of the claim that folk psychology is a theory. Patricia, like Paul, perceives the argument that first person ascriptions of psychological states are non-inferential and infallible, and are not, therefore, theoretical, as a principal objection to the claim. Her treatment of the argument is, as we shall see, largely concurrent with Paul's, although there are points of divergence.

The reason often given for taking self-ascriptions of psychological states to be non-inferential is that our knowledge of those states is immediate, which is to say that there is nothing mediating between the state ascribed and the judgement that one is in that state; nothing, that is, which would require that the judgement be made via an inference, as would be the case if a conceptual framework were interpositioned. For Patricia Churchland, however, this reasoning is fallacious. This is demonstrated, she argues, when we consider that perceptions of external objects can seem just as immediate as introspection of internal states and events, yet the fact that we are not always aware of the engagement of a conceptual framework when perceiving does not support the assumption that no such framework is engaged. Indeed, she urges, the assumption is a false one because 'complex information processing, pattern recognition, and conceptualizing certainly underlie ostensibly simple and "direct" perceptual judgements.' (ibid., p.306). If this is the case with seemingly immediate perception then we have no reason to believe that introspection does not involve processing in which a conceptual framework is engaged.

The argument will be unconvincing for those of us who might find it difficult to accept that we must assume much complex cognitive processing in accounting for such things as our ability to perceive that a tree occupying a similar area in one's visual field to that occupied by one's thumb is, in fact, a great deal larger than the thumb. The difficulty here is not due to the apparent immediacy of the perception but to the redundancy of the assumption. An exceptional circumstance could arise in which an explanation of the ability would be appropriate but it need be no more complex than the statement 'My thumb is only inches from my eyes whereas the tree is a couple of hundred yards away'. The requirement for a cognitive account of the perception results from the presupposition that perception must be explained as an inner process, and this is by no means unquestionable.

However, Patricia Churchland does not rely solely on an analogy between introspection and perception to demonstrate that recognition of one's own mental states entails the employment of a conceptual framework. Like Paul she adheres to the view that the psychological predicates used in the self-ascription of such states acquire their meaning from a theoretical network. The provenance of this view is the argumentation of W.V.O. Quine against the Logical Empiricist thesis that it is possible to define the terms of all meaningful epistemological claims using a language of pure observation devoid of any theoretical content (see Quine, 1951). The purity of the observation language is supposedly secured by the reference of its terms to sense-data because this reference relation is not mediated by a conceptual framework. An account of Quine's objections to this and connected empiricist theses, though interesting, would be somewhat tangential to our concerns at this point so I will shirk this task and allow Patricia Churchland to conclude that,

> 'The upshot of those arguments was that there is no independent observation language; that all observation terms are nodes in a conceptual framework and thus are interlaced with other observation terms and with theoretical terms; that all observation terms derive at least some of their meaning from the generalizations that embed them.'(Churchland, P.S., 1986, pp.306&307).

She goes on to say, 'Learning to apply mental predicates crucially involves learning appropriate generalizations that specify the conditions of correct application.'(ibid., p.307). These generalisations, such as 1) above, are to be understood as constituent of the folk theory and though, in a given case of self-ascription, the mental predicate may be applied non-inferentially, the background theory must already be in place in order to guide the speaker in his or her self-ascription.

This last point has relevance to a theme which Patricia Churchland lights upon and which I shall explore in subsequent chapters. The generalisations of folk psychology are to be understood as causal generalisations insofar as they assert nomic relations between states and events occupying a causal role. In the instance of 1) the person's states of wanting and believing are the causes of his performing *p*. However, if this is how we are to understand the generalisations then the consequence of Churchland's claim is that we are 'guided' in our use of mental predicates by causal laws rather than grammatical rules, Or, perhaps more correctly, acceptance of the thesis that folk psychology is indeed theoretical allows us to conclude that the rules governing the use of mental predicates are, in fact, causal laws. The conflation amounts to a surreptitious elimination of the distinction between the *normative* and the *causal*. Ultimately the conflation is essential to a science which aims to show how propositional attitudes (which bear normative relations to each other and to experience and behaviour) can be explained as either functional/computational, or neurophysiological states and processes.[3]

b) The Co-Evolutionary Approach

Having examined Patricia Churchland's statement, and defence, of the thesis that folk psychology constitutes a theoretical framework we can return to her claim that it is a candidate for reduction to neuroscientific theory.

As I suggested earlier, this seems to contradict Paul Churchland's antireductionist stance. Indeed, it appears contrary to the central tenet of Eliminativism which has it that folk psychology will be *replaced* rather than *reduced.* The contrariety becomes explicit when we consider that a prerequisite of candidacy for a reduction must surely be that the theory to be reduced employs categories with ontological status. After all, the outcome of such a reduction will be the identification of phenomena categorised in the reduced theory with phenomena categorised in the reducing theory, and this identification entails that the reduced categories are realized. For example, it is claimed that the theory of optics has been reduced to the theory of electromagnetic radiation and central to the claim is the identification of light with electromagnetic radiation. The reduction clearly affirms, rather than denies, the existence of the phenomenon of light. By analogy, if folk psychology is to be reduced to neuroscience then the phenomena described by the former must exist if there are to be the identity statements which are the mark of a successful intertheoretic reduction. Put another way the reduction of folk psychology to neuroscience seems to entail a Realist position with regard to mental phenomena and it is this position that Eliminativists, such as

[3] Churchland combats the Classical Computationalists' use of the distinction between the logical and the causal to outlaw the reduction of psychological to neurophysiological states. However, though the distinction is exploited by these theorists in their antireductionist arguments it is ultimately confounded by them in their quest for a computational model of cognition. This is a line of argument I shall pursue in chapter 6 when discussing the representation of linguistic rules in a physical system.

Paul Churchland, wish to discredit.

The apparent tension between Patricia and Paul Churchland's positions is relieved by the former's observation that,

> 'Reductions in science are only rarely smooth reductions with uncomplicated cross-theoretical identifications. More typically, they are bumpy and thus involve varying degrees of revision to the reduced science. Sometimes the correction required is so massive that the candidate theory is better described as having been displaced outright.' (Churchland, P.S., 1986, pp.310&311).

Patricia Churchland speculates that revision to the point of displacement may be the fate of folk psychology. She stresses that whether the cross-theoretical identifications will be forthcoming is an *empirical* rather than an *a priori* matter. Appreciation of this distinction,

> '...reveals the possibility that what will eventually reduce to neuroscience are generalizations of scientific psychology that have evolved a long way from the home "truths" of extant folk psychology.' (ibid., p.312).

So,

> 'what may eventually transpire, therefore, is a reduction of the evolved psychological theory, and this evolved theory may end up looking radically different from folk psychology - different even in its *categorial* profile'

and,

> 'if that is the direction taken by the co-evolution of psychology and neuroscience, future historians of science will see folk psychology as having been largely displaced rather than smoothly reduced to neurobiology' (ibid.).

It seems fair to suggest that although Patricia Churchland says that this scenario is purely speculative she is advocating a co-evolution of psychological and neuroscientific theory. Once we have accepted that folk psychology is a theory it follows that its generalizations are susceptible to empirical investigation. If those generalisations are found to be inaccurate or otherwise inadequate, then it might be that they can be improved by being measured against the findings of research in another theoretical field purporting to explain the same phenomena. Of course, for Churchland, that field would be neuroscience and her contention is that neuroscience and psychology need each other. Neuroscience needs psychology because, 'it needs high-level specifications of the input-output properties of the system' and psychology needs neuroscience because, 'it needs to know whether lower-level specifications bear out the initial input-output theory, where and how to revise input-output theory, and how to characterize processes at levels below the top'. (ibid., p.373).

In order for the co-evolution of the two theories to get under way an initial input-output theory is required and 'broadly speaking, folk psychology is that initial theory'. However,

> 'We have already gone beyond folk psychology, and as neuroscience and psychology co-evolve, the likelihood is that the initial theory will by inches be revised, lock, stock, and barrel.' (ibid., p.374).

The end result of the co-evolution of the two theories will be the '*establishing of points of reductive contact*' (ibid.).

In summary, what Churchland envisages is not the reduction of folk psychology to neuroscience but the development of a scientific psychology which will co-evolve with neuroscience until its categories, which may be quite different from those of its folk ancestor, can be reduced to—identified with—the categories that neuroscience employs, or will employ. Expressed in this way we can see that the form of reduction Patricia Churchland advocates is not outlawed by Paul's reprobation of reduction quoted at the beginning of this chapter. Although she does not explicitly prosecute the elimination of folk psychology in *Neurophilosophy*, Patricia Churchland does undermine its pretensions to explanatory adequacy and does suggest that revision of the theory will be required (see, for evidence of this, ibid., section 9.5).

c) Reasons to Reject the Theory

Whether psychology will ultimately reduce to neuroscience is, for Churchland, an empirical question and is not something that can be ruled out by a priori arguments. One line of a priori argument might be that psychological states, processes, and events have a subjective experiential character, or 'raw feel', which could never be reduced to the categories of an experimental science such as neurology. (This consideration is the source of the 'individuation' objections to the Type Identity Theory mentioned in the previous chapter.) Another line of a priori argument is that cognition is fundamentally representational and the relations that exist between representations, and between representations and behaviour, are logical and, therefore, inaccessible to explanations proffered by neuroscience.

The second line of argument indicates that Churchland's case for reductionism would be made easier if she could rid psychology of the contents of propositional attitudes, for it is these which make the relations between mental states, stimuli, and behaviour rational and/or logical. It is here that her co-evolutionary brand of reduction becomes apposite. If psychological theory, informed by advances in neuroscience, evolves into a theory which does not recognise the propositional (or sentential, as Churchland has it) attitudes in its ontology, then the antireductionist argument evaporates. In *Neurophilosophy* Churchland sets out to expedite the departure of sentential attitude psychology by presenting three problems which she does not think it will overcome.

The first problem is dubbed 'The Infralinguistic Catastrophe' and is a re-working of Paul Churchland's sixth reason for the elimination of folk psychology as given above. Patricia Churchland's version of the argument runs from an observation, that non-linguistic creatures, such as deaf mutes, chimpanzees, octopi, and babies, can display intelligent behaviour, to a conclusion that this falsifies the thesis that intelligence requires computations ranging over linguistic structures (ibid., p.389). It should be noted that although the intelligent behaviour attributed to babies is language learning, this attribution need not conflict with the first premiss of Paul's version of the argument which assumes that newborns do not have propositional attitudes. Note also that the argument has force against the founding of cognitive theories on the folk psychological taxonomy but not against the everyday, non-scientific, use of psychological terms.

The second problem for sentential attitude psychology is that of 'Tacit Knowledge'. Bearing in mind that Fodor maintained that to have a propositional attitude is to be in a computational relation to a representation, this problem leads to the postulation of an infinite store of sentential representations. This is because a person can be attributed with an infinite number of tacit beliefs. To adapt Churchland's example (ibid., p.390), we might ask of Smith whether he believes his horse is less than three metres tall. Since his answer would be yes we may attribute him with the belief and hence the representation of the sentence 'My horse is less than three metres tall'. We may also ask him whether he believes that, for every number, *n*, greater than three, his horse is less than *n* metres tall. His affirmation of the belief will imply that he tacitly holds the beliefs that his horse is less than four metres tall, less than five metres, less than six, and so on. That is, he will have an infinite set of tacit beliefs about his horse's height, and we have not even asked him about his cat. Indeed, there is an infinity of matters on which Smith may have an infinite number of tacit beliefs and this makes talk of belief *storage* extremely problematical.

Churchland blocks the easy way out for the sentential theory—the denial that we have tacit beliefs—by pointing out that though a belief has not been explicitly entertained it may still have a causal role in ones behaviour and therefore can not be denied an instantiation. To illustrate her thoughts, I may never express or even consider my belief that the chair I am sitting on will hold my weight yet I may be said to act on the belief and if, for some suitable reason, I was asked if I hold such a belief I would certainly answer yes.

An alternative way of avoiding the problem is to take Daniel Dennett's suggestion and assume there is an extrapolator-deducer mechanism attached to a core library of sentential representations. Of this Dennett says;

> 'It has the capacity to extract axioms from the core when the situation demands it, and deduce further consequences. To do this, it needs to have an information store of its own, containing information about what items it would be appropriate at any time to retrieve from the core...'(Dennett, 1975, p.45).

However, this suggestion gains no explanatory ground and leads to an infinite regress as Dennett demonstrates:

> 'Now how will the extrapolator-deducer mechanism store its information? In its own core library of brain-writing sentences? If it has a core library, it will also need an extrapolator-deducer mechanism to act as librarian, and what of *its* information store?'(ibid., pp.45&46).

So a mechanism which would generate implied beliefs from a store of explicit axiomatic belief representations would require a further store of representations to inform it as to which axioms are relevant to the situation in which the implied beliefs will be appropriate. A similar mechanism would be required to extrapolate from this store, and that mechanism would require its own store, hence the infinite regression.

This objection certainly gives reason to doubt the coherence of the account of tacit knowledge which makes appeal to stores of beliefs represented as sentences. Churchland concludes;

> 'Thus, the supposition that the knowledge store is a sentence (belief) store comes to be regarded as untenable. Abandoning that supposition, we can try instead the idea that tacit knowledge is not (mainly not) a corpus of tacit beliefs....What is stored is

> generally something else, something that may be verbally encoded on demand, but need not be verbally encoded to be cognitively engaged....Other sorts of representational structures will need to be postulated; hence the interest in nonsentential representations such as prototypes, images, and frames.'(ibid., p.392).

Whether the objections warrant this conclusion is questionable. The fact that *any* postulation of internal representations in the explanation of cognition invites the use of the infinite regress argument, as a *reductio ad absurdum*, is something I shall argue for in chapter 4.

The last of the problems for the sentential model of cognition is that of 'Knowledge Access'. This can be expressed an a problem in explaining how an information processing system ascertains which part of its sentence store will be relevant to the situation in which it finds itself. It is a difficulty which has been faced by researchers in AI when they have tried to produce models of systems which can be programmed to solve problems. Once again Churchland refers to Dennett, this time for an illustration of the problem, and I shall follow suit.

We are to imagine a robot, R1, is given the task of removing its spare battery from a locked room in which a time bomb is due to go off. It finds the key, unlocks the door, sees the battery on a trolley and conjectures that a procedure called PULLOUT (TROLLEY, ROOM) will get the battery out of the room. R1 has also seen that the bomb is on the trolley but failed to infer the consequences of this fact.

Having picked up the pieces of R1 the designers set about producing a robot which will deduce both the intended and unintended consequences of its action, and this they call R1D1. The task remains the same and R1D1, like its predecessor, opts for the PULLOUT (TROLLEY, ROOM) procedure but first must run through the implications of this action, in accordance with its program. Having deduced that removing the trolley from the room will not alter the colour of the walls, it begins to prove that the number of wheels on the trolley is less than the number of revolutions through which they will turn on their way out of the room. It is during this deduction that the bomb explodes.

This time the designers realise that what they need is a program which selects the relevant implications for consideration and ignores the irrelevant ones. The robot R2D1 is given just such a program and the same task as before. However, this time the robot does not even get into the room because it must postpone action until it has ignored the thousands of irrelevant implications that action might have. The delay is too long and, as we know, time bombs and tide wait for no one (Dennett, 1984, pp.129&130).

In these scenarios each robot is able to gather information from its environment which it stores in sentential form (presumably using formulae of its machine code). The problem begins to become apparent when we note the abundance of information that can be adduced from even a simplistic environment. The problem becomes similar to that of tacit knowledge because the sentential representation of a fact will carry with it a multitude of implied facts and these will multiply as soon as the environment is changed either by the robot or by the designers.

The point of Dennett's illustration and, in particular, the point of placing a time bomb in the room, is that if the robot can not act quickly in the situation by employing its sentential representation system for problem-solving, then there is good reason to reject the claim that humans employ such a system since most people would be capable of devising a way of retrieving the battery into safety in very little time. So whether Churchland is correct to see this problem as striking a death knell for sentential attitude psychology will depend on whether it can be solved, theoretically at least, by those in the AI field who use sentential models for problem-solving systems.

One approach to the solution of the problem has been to employ the, so-called, attention-focusing power of stereotypes. Dennett explains;

> 'The inspiring insight here is the idea that all of life's experiences, for all their variety, boil down to variations on a manageable number of stereotypic themes, paradigmatic scenarios—"frames" in Minsky's terms...'(ibid., p.144).

(It is worth noting that the general problem of knowledge access is sometimes called the 'Frame Problem' by AI researchers though this title is also used to refer to a more localised problem regarding the manner of representation of information. Dennett uses the title to name the more general problem (see ibid., pp.130&131)). The strength of this approach is that it seems to capture a central aspect of human rationality which facilitates rapid problem-solving. When facing a task we are able to ignore much of the irrelevant information we might have about our environment, as well as its implications, by making a *ceteris paribus* assumption. To illustrate, when given the description of the robot's task as that of removing the spare battery from a locked room containing a time bomb, we humans can plan our action swiftly because we assume normality in the situation. We do not consider, for example, that the walls of the room will collapse as soon as we have removed the battery, or that the floor has a screed of newly laid quick-drying cement, and so on. Instead we assume that, beyond the features given in the description, all things are equal and it is this which allows us, but not R2D1, to 'jump to conclusions'.

So the knowledge access problem challenges the designer of sentential models to produce a system which will have the ability to ignore many features of its environment when devising methods for performing tasks. Dennett's worry is that these models are in danger of postulating what he calls 'cognitive wheels'. For our purposes we need only consider Dennett's explanation of cognitive wheels as

> 'any design proposal in cognitive theory (at any level from the purest semantic level to the most concrete level of 'wiring diagrams' of the neurons) that is profoundly unbiological, however wizardly and elegant it is as a bit of technology'(ibid., p.147).

So even if a robot could be produced which could solve practical problems in something like the time span that most humans can, the likelihood is that its wiring and programme would not resemble anything that could be found in a human's neurophysiology.

Churchland interprets Dennett's scepticism as warranting the conclusion that,

> 'the more we try to solve the robot's problem of sensible behavior, the more it becomes clear that *our* behavior is not guided by explicit sentential instructions in our store of knowledge. Specifying the knowledge store in sentences is a losing strategy.' (Churchland, P.S., 1986, p.394).

Dennett does acknowledge that dispensing with the predicate-calculus format used for representing 'propositions believed' (the inverted commas are used by Dennett) is a 'tempting suggestion' but does not assert that the use of this format is a 'losing strategy'.[4]

In fact, not one of the three problems Churchland presses into service as refutations of sentential attitude psychology has been recognised as such by adherents to this approach. As Patricia Kitcher puts it;

> 'No one in the sententialist camp is going to abandon the position on the basis of Churchland's claim that there are problems with babies, animals, closure, and frames. Sententialists fully acknowledge the closure and frame problems and are actively working on solutions. Further, they do not believe that babies and animals talk; they believe that the best way to model higher mental life is in terms of inner representational states that have features such as constituent structure in common with natural languages. In effect, Churchland's objection is that it is not obvious how

[4] Towards the end of the paper, Dennett suggests that the frame problem might be restricted to 'the conceptual scheme engendered by the serial processing von Neumann architecture of the computers used to date in AI'. He then acknowledges the development of 'fast parallel processors' which may bring conceptual innovations pointing to a path around the problem (Dennett, 1984, p.149). In addition, 'Since brains are surely massive parallel processors, it is tempting to suppose that the concepts engendered by such new hardware will be more readily adaptable for realistic psychological modelling' (ibid.).

So parallel distributive processing, or Connectionist, architectures may not only solve the frame problem but may also do it in the way that human brains do. However, appearance to the contrary, this does not bode well for the Classical Computationalist/sentential attitude psychologist because, as we shall see in the next chapter, Connectionism seems to favour Eliminativism.

> the sentential model is going to handle these phenomena or that the model does not now have a satisfactory account of them.'(Kitcher, 1996, p.52).[5]

Since Kitcher is writing ten years after Churchland it does not seem that the problems were quite as threatening to propositional attitude psychology as she would have had us believe. For my own part, however, I believe that the problems do indicate an inherent incoherence in explanations of cognition by appeal to internal processes. For it is solely as a result of this appeal that the need arises for positing the subconscious manipulation of internally represented information. Churchland's conclusion that the information is not represented in sentential but in some other form may be no less incoherent, for while the sentential model becomes burdened by an unmanageable and, particularly in the case of the pre-linguistic, unlikely knowledge store, it seems that the non-sentential approach has the burden of accounting for the contents of cognitive states and processes. That is, if, as seems to be the case, Churchland wants such states and processes to remain among the explananda of scientific psychology, then she will need an account of how facts about the world are presented to the cognizing subject. Among the rival accounts, the sentential one has the greatest plausibility for it can happily attribute internal structures with propositional content.

It is here that Churchland's Eliminativism becomes apposite. If we were to remove the notion of propositional content from cognitive process explanation then the problems we have been considering would not arise. That this would force into obsolescence rational explanation couched in the vocabulary of the propositional attitudes is acceptable enough for Churchland. She cites the work of Stephen Stich as a source of evidence that 'the folk psychological categories of belief and desire are bound to fragment at the hands of science'(ibid., p.382). Stich's contribution was to show that, 'belief ascription is context-relative, and depending on interests, aims, and sundry other considerations, different criteria

[5] The problem of 'closure' is the same as that of Tacit Knowledge just considered.

are variously used to specify the content of a given mental state.'(ibid.). Thus the content of a belief will not be determined solely, if at all, by any form of internal representation but by criteria based on extraneous factors. Churchland takes this as further justification for viewing folk psychology as a false theory. She writes,

> 'If Stich is right, and I think he is, then it is already clear that the propositional attitudes qua folk psychological categories, are coming apart. Therefore, when antireductionists parade these categories in all their folk psychological regalia as irreducible, the irony is that it is their lack of empirical integrity that prevents their reduction.' (ibid., p.383).

We shall be looking at Stich's work in the next section. It is worth noting here, however, that though his quasi-Wittgensteinian observation is most certainly damaging to the sentential account of cognitive processing, it is only when this is confused with folk psychology (in the sense of everyday usage of psychological predicates) that the observation and the aforementioned problems give Eliminativism appeal.

At times, Churchland clearly believes that folk psychology, as a theory, either entails or contains sentential attitude psychology's postulation of sentential representations (see, for example, ibid., p.385 and pp.396&397). Her confusion no doubt arises from an inability to appreciate that the ascription of propositional attitudes does not require the adoption of a Realistic stance regarding them. (Perhaps it is the very phrase 'propositional attitude' that invites the confusion because it encourages us to postulate the existence of a sentential structure to which we might be related). Thus, the fact that propositional attitude terms are an integral part of the vocabulary of folk psychology does not entail that it is committed to Realism which is why it must be differentiated from a sentential attitude psychology of the form defended by Fodor.[6]

[6] In saying that folk psychology, *qua* everyday psychological language, is not committed to Realism I do not mean that it is ontologically neutral, for Realism is itself neutral in this respect (which is why both Descartes and Fodor can be Realists). Rather what I mean is that folk psychology is not committed to the view that psychological ascriptions are intended to refer to states, processes, and events, which can be identified independently of their behavioural manifestations. On this understanding there is

Sentential attitude psychology does require a Realistic stance regarding propositional attitudes but everyday psychological discourse does not. As Kitcher notes, the opposite would be true only 'if it could be shown that the average person believes that mental states are inner representational propositional states' (Kitcher, 1996, p.79 note 2). This would imply a general acceptance and understanding of the Representational Theory of Mind, in sentential form, and that is not the case. If this is a postulate of the 'theory' theory of folk psychology, then this reflects badly on the theory.[7]

Returning to the exposition, we might understand Eliminativism as an attempt to naturalize the notion of mental content by replacing (as opposed to identifying) it, in explanations of the ætiology of cognition, with descriptions of purely physical causes. This proposed replacement, as we have seen, may be prompted by the rejection of the claim that the brain has structures that bear mental content—the claim if the Classical Computationalist, and sentential attitude psychologist. In the next section I will survey some further reasons for rejecting the claim, of which some support the Eliminativist thesis.

3 STEPHEN STICH

a) The Syntactic Theory of Mind

According to some of the philosophers who contribute to cognitive science Connectionism and Eliminativism are complementary positions. Stephen Stich takes this view, as we shall see in the next chapter. However, before looking at Stich's more recent Connectionist leanings it would be helpful, in our circumscription of the philosophical issues pertaining to cognitive

no ontology of propositional attitudes.

[7] As a point of interest, Kitcher finds three distinct uses of the expression 'folk psychology' in Churchland's writing. It is used to refer to a) 'the inchoate psychological theory held by ordinary people', b) 'the view that [philosophically relevant] mental phenomena can best be understood by positing the existence of and transformations among propositional or sentence-like inner representations', and c) 'those parts of traditional psychology that can not be grounded in neurophysiology' which are, consequently, to be eliminated (Kitcher, 1996, pp.51-53).

Of particular interest is the failure of all of these uses to capture the way in which psychological terms are used by 'folk'.

science, to place his earlier work in relation to that of the Churchland's. In doing so we shall add further background to the discussion of the notion of content arising later.

In *From Folk Psychology to Cognitive Science* Stich attacks Fodor's Classical Computationalism and, in particular, his Representational Theory of Mind. I say 'in particular' because he can be seen as endorsing other aspects of Fodor's theory of cognition. To appreciate this we need to look, albeit briefly, at Stich's alternative model for cognitive science, namely the Syntactic Theory of Mind (hereafter, the STM).

Stich explains that,

> 'The basic idea of the STM is that the cognitive states whose interaction is (in part) responsible for behavior can be systematically mapped to abstract syntactic objects in such a way that causal interactions among cognitive states, as well as causal links with stimuli and behavioral events can be described in terms of the syntactic properties and relations of the abstract objects to which the cognitive states are mapped. More briefly, the idea is that causal relations among cognitive states mirror formal relations among syntactic objects.' (ibid., p.149).

As he admits, this is not a cognitive theory but, like the Representational Theory of Mind, it is a 'paradigm for cognitive theorizing' (ibid.). The production of a cognitive theory on the STM model demands that the theorist fulfils three main tasks as follows:

1) He needs to specify the class of syntactic object types in such a way that a formal or syntactic structure is assigned to each. In doing so he will need to employ a grammar or set of formulation rules detailing the generation of complex syntactic objects from a finite set of primitives. This is because of the fact that, commonly, the class of object types will be infinite (ibid., p.150).

2) He needs to hypothesise firstly, that for each organism covered by the theory, there exists a set of state types the tokens of which have a causal role in the production of behaviour and,

secondly, that there is a mapping from these state types to syntactic objects in the specified class.

3) He needs to specify the theory's generalisations. To accord with the STM model the generalizations detailing causal relations among the hypothesized neurological states, and between those states, stimuli, and behaviour, must be specified not by adverting directly to their essential neurological types, but indirectly via the formal relations among the syntactic objects to which the neurological types are mapped (ibid., p.151). In other words, the generalizations will range explicitly over syntactic objects and only implicitly over their neural realizations.

Stich does not presume to foretell the form of the syntactic objects a STM would specify but, for ease of comparison with content-based theories, he suggests that the objects may constitute 'a class of sentences or, better, *well-formed formulas* (wffs)' (ibid., p.153). The formulae would have a underlying syntax such as (for the sake of illustration) that of first-order quantification theory. Although Stich adds a lot more detail to his account of the STM, we now have just enough outline to superimpose Stich's paradigm onto Fodor's in order to see how much is overlap and how much unique to each.

Fodor, as we have seen, maintains that mental states are relational and that among the relata are mental representations (that is, formulae in the LOT). These representations have both formal ('roughly' syntactic) and semantic properties—examples of the latter being properties of reference, meaning, and truth (see Fodor, 1980, p.227). Further, mental states inherit their semantic properties from those of the representations functioning as their objects, and it is the formal properties of the representations which dictate their causal roles and, therefore, assist in determining which mental state (that is, which propositional attitude) the representation contributes to (Fodor, 1975, p.198 and 1981, p.26).

Thus, *prima facie*, it would seem that Stich and Fodor concur firstly in maintaining that formulae are constitutive of mental states, and secondly in citing the formal properties of formulae as the determinants of the causal roles of those states. When we add to this the STM's apparent requirement of a generative grammar (see 1) above) the function of which parallels that postulated by Fodor, it becomes tempting to see the STM as a variant of the LOT hypothesis.[8] However, before succumbing to the temptation one should be aware of the discord between the two positions.

The models proposed by Stich and Fodor may overlap in their functional characterisation of mental states as relations to mental sentences but while the STM would 'treat mental states as relations to purely syntactic mental *sentence* tokens' (Stich, 1983, p.9, my italics) Fodor's LOT hypothesis imbues these sentences with semantic properties in the form of propositional content. The importance of this difference becomes apparent when it is appreciated that while the generalisations envisaged as issuing from the LOT model will range over propositional attitude states those from the STM will not be couched in the 'content-ascribing language of folk psychology' (ibid., p.8). It is this aspect of Stich's theorising which earns him his Eliminativist colours.

In subsequent chapters I will give a fuller account of some of the views about content that have worried Fodor and other Classical Computationalists. At present, however, I will continue by exploring Stich's position beginning with his stance as it relates to Eliminativism.

[8] In addition, both Stich and Fodor make the Functionalist provision for the multiple instantiation of cognitive states which is indicative of a token physicalism. We already know of Fodor's token physicalism which received attention in the first chapter. Stich shares this stance on the ontology of mental states for he says, 'one way of construing the view I have been sketching is as a sort of token identity theory for beliefs' (Stich, 1983, p.223). As we shall see shortly Stich also seems to throw in his lot with the Eliminativists who, of course, deny the reality of folk psychological states and, thereby, preclude their appearance in identity relations. However, Stich attempts to avoid contradiction by maintaining that though cognitive science will not invoke folk psychological concepts in its explanation of behaviour, it will postulate state tokens which can be described as the belief or desire that *p* (ibid., p.224).

b) The Folk Theory

Despite showing strong Eliminative tendencies in his denial of a place for folk psychology in cognitive science, Stich begins the tenth chapter of *From Folk Psychology to Cognitive Science* with a rebuttal of some of the Churchlands' arguments for a similar denial. Stich finds faults in what he calls the '"degenerating research program" argument'. This argument covers the first three of the six reasons for eliminating folk psychological theory that I attributed to Paul Churchland in section 1c) of this chapter. In brief, the reasons were as follows:

1) The explanatory power of folk psychology is severely limited for it lacks an adequate account of such things as mental illness, memory, creative imagination, and sleep.

2) As a theory it has degenerated, as the decline of animism attests, and stagnated since there has been little advancement in the last two millennia.

3) It fails to cohere with the rest of the physical sciences because of its psychological categories which resist reduction to, or synthesis with, those sciences.

Stich offers objections to all three reasons. The most significant objection, for our purposes, can be levelled at 1) and 2) and makes the point that 'folk psychology is not properly viewed *merely* as a crude explanatory, scientific theory, since the terms of folk psychology have more work to do than do scientific terms' (ibid., p.212). Stich attributes the observation to Kathleen Wilkes who has argued that although folk psychology (or 'common-sense psychology' as she prefers to call it) shares with the sciences the tasks of explaining, describing and predicting, the former has a multitude of other tasks which it does not share with the latter (Wilkes, 1981, pp.149&150). Wilkes gives as examples of such tasks warning, threatening, assessing, applauding, praising, blaming, discouraging, urging, wheedling, sneering, hinting, implying, and insulting, and summarises by saying that 'the conceptual apparatus of common-sense psychology stands to that of scientific psychology as a multi-purpose tool stands to a spanner' (ibid.).

Wilkes's argument is that common-sense psychology is not a theory 'in any useful sense of the term "theory"' but Stich finds this too strong a claim. He maintains that in using folk psychological concepts to explain and predict the behaviour of others we presuppose 'rough and ready laws which detail the dynamics of belief and desire formation and connect these states to behavior'. The laws can be 'teased out and made explicit' and so 'collectively they surely count as a commonsense theory' (Stich, ibid.). Wilkes has more to say on the subject (see note 10 below) but for Stich the weaker claim (that it is a 'multi-purpose tool') suffices to excuse the failure of folk psychology to progress noticeably in the past two millennia or to fill explanatory lacunæ given that its concepts 'served well in their non-protoscientific roles'. After all, since its assumptions were not recognised as protoscientific folk psychology could not be expected to evolve better ones (ibid., p.213).

We might add that Stich dissents from the view that folk psychology is a stagnating theory, since it has only recently become an experimental discipline, and he suggests that it is indispensable to the social sciences which excuses its incompatibility with the physical sciences. Given that he is in disagreement with the arch-Eliminativist, Paul Churchland, the question arises once again as to Stich's relationship to Eliminativism.

Paul Churchland has continually claimed that folk psychology is a false descriptive and explanatory conceptual framework due for elimination and replacement by a matured neuroscience or, to bring us more up to date, by 'a conceptual framework of sufficient richness and integrity that you will be able to reconceive at least some of your own mental life in explicitly neurocomputational terms' (Churchland, P.M., 1995, p.19). Stich, in acknowledging heterogeneity in the functions of folk psychology and its indispensability in certain areas of systematic explanation, could not endorse Churchland's proposal of a global eradication of the conceptual framework. That said, Stich certainly does argue for the more modest elimination of folk psychology from the area of our concern, namely cognitive science.[9]

c) Reasons to Reject the Theory

It is interesting that Stich's reasons, for doubting the suitability of the folk framework for cognitive theorising arise, like those of the Churchlands, from a conception of folk psychology as a theory which generalizes over the relations between mental states, and between mental states, environmental conditions and behaviour. Furthermore, some of the reasons for doubt are based on the assumption that folk psychology entails a commitment to a Sententialist account of the propositional attitudes. Thus, Stich gives as one reason to banish folk psychology from cognitive science the context relativity of folk psychological concepts (ibid., pp.217&218) and offers, as another, a variation on the 'infralinguistic' problem (Stich, 1983, pp.135-148).

The first reason has Wittgensteinian provenance and shows that folk psychological predicates cannot be projected into the nomic statements required by a science of cognition.[10] The second reason rests on the assumption that, in attributing beliefs, the folk are attributing mental sentences or, at least, discrete mental states to themselves and to non-humans. This assumption also lies behind another charge Stich makes against folk psychology, to wit, that the functional organisation of the mind presupposed by the folk theory requires a certain degree of *modularity* in the organisation of the belief or memory store. He explains that,

[9] On the other hand, by the end of the book, Stich seems to expect that folk psychology will be found to be a false theory in which case 'our age old conception of the universe within will crumble just as certainly as the venerable conception of the external universe crumbled during the Renaissance' (Stich, 1983, p.246). If we take this as implying that folk psychological concepts will disappear not just from cognitive science but also from everyday usage then Stich's position is hardly distinguishable from Paul Churchland's.

[10] Wilkes also makes the point that common sense psychology does not offer a 'systematic taxonomy' for the description of 'repeatable' observations which are 'context-*transcendent*' (1991, p.174). However, for her this gives grounds to reject the view that common sense psychology is a theory 'in any substantial sense of that term' (ibid., p.170). In my view Wilkes is correct.

> 'A belief or memory storage system is *modular* to the extent that *there is some more of less isolatable part of the system which plays (or would play) the central role in a typical causal history leading to the utterance of a sentence.*' (ibid., pp.237-238, Stich's italics).[11]

Such an assumption is clearly made by a sentential theorist who views memory as a list of sentence-like structures, with each one corresponding to a separate belief and helping to cause a self-ascription of that belief.[12] Stich cites John McCarthy as an example of such a theorist but the assumption should be attributed to Fodor also and, indeed, anyone (excepting an Epiphenomenalist) who might be described as a Realist regarding the propositional attitudes (although some may prefer a non-sentential model).[13] To understand Stich's attitude to this modularity assumption we need to address two questions. Firstly, what would the implication be, for belief ascriptions, if the assumption proved false? Secondly, what reason might we have to question it?

Stich views the ascription of a belief to someone, roughly, as a statement to the effect that the person has *something* similar to what we would have were we to utter a true sentence expressing the same belief. According to sentential attitude theories that *something* is a content sentence which contributes to the belief state, and is causally efficacious in producing the utterance. Therefore, the system in which the belief is found will be modular by Stich's definition of that term. Now, if the modularity assumption is false, then 'belief ascriptions will typically lack a truth value' because, as Stich sees it, such ascriptions have the form 'X is in a

[11] Stich's use of 'modularity' should not be confused with Fodor's in his book *Modularity of Mind* (Fodor, 1983). Although he is reluctant to define modularity in this book (see ibid., p.37), we can glean from what Fodor does say that a psychological module is a system with its own information store and that the operations it performs have access to this store alone.

[12] Note that being sentential is sufficient but not necessary for a model's being 'modular' on Stich's definition of that term, for isolatable parts of the system may cause utterances of sentences without having propositional content themselves. In the next chapter we will see that, as a polemical device, modularity was to acquire semantic properties to add to the causal ones in its later incarnation as *propositional* modularity.

[13] In fact, by the lights of his definition of modularity Stich himself makes the assumption because, for him, B-states are the causal antecedents of utterances of some sentences (Stich, 1983, p.154) and that they are isolatable is presupposed by the stipulation that they can be mapped to well-formed formulas. Stich himself alludes to this at a later date when he says that 'neither Fodor's account of cognitive theorizing nor my syntactic account mesh comfortably with the connectionist paradigm' (Stich, 1991, p.252 note 4). However, (as we shall see shortly) the reasons Stich cites for rejecting the modularity assumption attach to the content of the cognitive states which are supposed to cause utterances, and his B-states do not have semantic content (though, by virtue of the mapping, they do have a syntax).

belief state similar to *the belief state which would play the central causal role if our utterance of the content sentence had had a typical causal history*' and, therefore, will be invoking a definite description (here italicized) which fails to denote (ibid.). That is, if the assumption is false, then there are no belief states. Thus, because of his Simulation Theory of belief ascription, Stich maintains that the validity of our practice of belief ascription will depend on the truth of the modularity assumption. So, why might we question the assumption?

Stich briefly mentions the problems arising from the retrieval of stored information relevant to a discourse (ibid., p.239), but since these are aspects of the 'frame problem' we have already encountered I shall pass over them. It will suffice to say that they are undoubtedly troublesome for modular theories of cognition such as are found in sentential attitude psychology, as we acknowledged in the previous section.

Given that there are problems with the modularity assumption (especially in so far as it presupposes the folk psychological taxonomy), the existence of alternative models of cognition which evade those problems would seem a good reason for, at least provisionally, abandoning it. Stich cites the work of Winograd and Minsky as a source of such alternative models and regarding the latter he says,

> 'On the picture Minsky suggests, none of the distinct units or parts of the mental model "have meanings in themselves" and thus none can be identified with individual beliefs, desires, etc. Modularity...is violated in a radical way since meaning or content emerges only from "great webs of structure" and no natural part of the system can be correlated with "explicit" or verbally expressible beliefs.' (ibid., p.241 with quoted expressions from Minsky, 1981, p.100).

In the previous section I mentioned Dennett's speculation that the frame problem could be avoided by systems simulating non-deductive inference which can make *ceteris paribus* assumptions about their environment. Stich makes a similar speculation by quoting Minsky's suggestion that,

> 'the strategy of complete separation of specific knowledge from general rules of inference is much too radical. We need more direct ways for linking fragments of knowledge to advice about how they are to be used.' (Minsky, 1981a, p.127).

Models of aspects of cognition with the properties favourably mentioned by Dennett, Minsky, and Stich, were starting to appear in the early eighties with the emergence of Parallel Distributed Processing which yielded Connectionist architectures. If we keep in mind the fact that Stich's STM abjures propositional attitudes, and therefore does not need to postulate content bearing structures, we will see why Connectionist models might appeal to him.

CHAPTER THREE

CONNECTIONISM

1 CONNECTIONISM AND ELIMINATIVISM

a) Propositional Modularity

In 'Connectionism, Eliminativism, and the Future of Folk Psychology' by Ramsey, Stich, and Garon, the conditional claim is made that, '*if* Connectionist hypotheses of the sort we will sketch turn out to be right, so too will eliminativism about propositional attitudes' (Ramsey et al, 1990, p.312). The hypotheses referred to here contribute to models of belief and memory to which we will turn shortly. Of course, as the authors acknowledge, those who believe that Eliminativism is wrong can conclude, by *modus tollens*, that the favoured Connectionist hypotheses are, therefore, incapable of modelling human cognitive activity.

The premisses for the argument that the Connectionist models, provided by Ramsey et al, entail elimination of propositional attitudes are, firstly, that 'common sense psychology is committed to propositional modularity' and, secondly, that the models 'are *not* readily compatible with propositional modularity' (ibid., p.320). What we require, then, is an account of propositional modularity.

Common sense psychology is committed to propositional modularity because it assumes,

> 'that propositional attitudes are *functionally discrete, semantically interpretable*, states that play a *causal role* in the production of other propositional attitudes, and ultimately in the production of behavior.' (ibid., p.315).

If common sense psychology is to be understood as familiar discourse about people's beliefs, wishes and the like then this claim is incredible. However, if we suspend disbelief and accept, for the moment, the view that participation in common sense psychology presupposes tacit adoption of a theory about the ætiology of psychological states and processes and their concomitant behaviour, then we can imagine why Ramsey et al feel justified in making the claim.[1] That said, it would be helpful to provide more detail.

The claim is that common sense psychology assumes propositional attitudes are;

1) *functionally discrete* because 'it typically makes perfectly good sense to claim that a person has acquired (or lost) a single memory or belief', although this is not to deny that having a particular belief may presuppose a network of related beliefs. We are given the example of a belief that the car keys are hidden in the refrigerator by way of illustration (ibid., p.316);

2) *semantically interpretable* because common sense psychology takes propositional attitudes to have semantic properties. For example, 'when people who speak English say "There is a cat in the yard", they typically believe *that there is a cat in the yard*'. We should note that generalisations such as this are 'couched in term of the *semantic* properties of the attitudes' and that common sense psychology treats the predicates expressing these properties as of the sort that are used in nomological, or law-like, generalisations. Thus, we are told, 'it is in virtue of being the belief *that p* that a given belief has a given effect or cause' (ibid., p.316);

3) states with a *causal role* because 'in common sense psychology, behavior is often explained by appeal to certain of the agent's beliefs and desires' (ibid., p.317). To concoct an example, it is my belief that the car keys are in the refrigerator combined with my desire to find them which causes me to open it and remove them. In addition it makes sense to say, of a pair of occurrent beliefs, which particular belief caused the agent to act, thus demonstrating that beliefs are causally discrete.

I have already suggested that Fodor can be accredited with the modularity assumption explicated at the end of the previous chapter. The account of propositional modularity is, of

[1] I will postpone criticism of the view until the last chapter.

course, a closer approximation to the characterisation of propositional attitudes one finds in Fodor's writing because, for him, mental representations have semantic as well as causal properties. Again, if it is the case that, Epiphenomenalism notwithstanding, to be a Realist regarding propositional attitudes is to be committed to propositional modularity, then if the correct model of human cognitive activity does not display propositional modularity it follows that Realism is false because there is nothing in the model corresponding to belief states, desire states, thoughts, and so on. It is not difficult to see how the inference is made from the falsity of Realism to the truth of Eliminativism. Admittedly, Ramsey et al do not presume to know of the correct model yet—as we have noted, theirs is only a conditional claim. However, they do present an outline of a Connectionist model of memory and appreciating why it does not have propositional modularity will be facilitated by the general characterisation of Connectionism assented to by Ramsey et al (ibid., pp.320&321).

According to Paul Smolensky in 'On the Proper Treatment of Connectionism' (Smolensky, 1988),

> 'Connectionist models are large networks of simple, parallel computing elements, each of which carries a numerical *activation value* which it computes from neighbouring elements in the network, using some simple numerical formula. The network elements or *units* influence each other's values through connections that carry a numerical strength or *weight*. The influence of unit *i* on unit *j* is the activation value of unit *i* times the strength of the connection of *i* to *j*. Thus, if a unit has a positive activation value, its influence on a neighbor's value is positive if its weight to that neighbor is positive, and negative if the weight is negative...
>
> In a typical connectionist model, input to the system is provided by imposing activation values on the *input units* of the network; these numerical values represent some encoding, or *representation*, of the unit. The activation on the input units propagates along the connections until some set of activation values emerges on the *output units*; these activation values encode the output the system has computed from

> the input. In between the input and output units there may be other units, often called *hidden units*, that participate in representing neither the input nor the output.
>
> The computation performed by the network in transforming the input pattern of activity to the output pattern depends on the set of connection strengths; these weights are usually regarded as encoding the system's knowledge. In this sense, the connection strengths play the role of the program in a conventional computer. Much of the allure of the connectionist approach is that many connectionist networks *program themselves*, that is, they have autonomous procedures for tuning their weights to eventually perform some specific computation. Such learning procedures often depend on training in which the network is presented with sample input/output pairs from the function it is supposed to compute. In learning networks with hidden units, the network itself "decides" what computations the hidden units will perform; because these units represent neither inputs nor outputs, they are never "told" what their values should be, even during training.' (Smolensky, 1988, pp.28&29).

Ramsey et al add to Smolensky's characterisation the point that,

> 'in many connectionist models the hidden units and the output units are assigned a numerical "bias" which is added into the calculation determining the unit's activation level. The learning procedures for such networks typically set both the connection strengths and the biases. Thus in these networks the system's knowledge is usually regarded as encoded in *both* the connection strengths and the biases.' (Ramsey et al, 1990, p.321).

Their Connectionist model of memory consisted of a three tiered feed-forward network incorporating sixteen input units, four hidden units and one output unit. Outputs close to one were interpreted as 'true', while those close to nought were interpreted as 'false'. The network was 'trained up' using a back propagation algorithm which set the connection weights and biases. The network is said to have stored information about the truth or falsity of sixteen propositions because 'when any one of these propositions is presented to the network it correctly judges whether the proposition is true or false' (ibid., p.325). What we require, then, is an explanation of why this network is incompatible with propositional modularity.

A computational model of cognition which has propositional modularity might be expected to represent a feature of the environment via a grouping of symbolic elements which would have a unique identification with a propositional content.[2] However, Ramsey et all point out that,

> 'in models where the weights and biases have been tuned by learning algorithms it is often not the case that any single unit or any small collection of units will end up representing a specific feature of the environment in any straightforward way...'

and although,

> 'it is often plausible to view such networks as collectively or holistically encoding a set of propositions... none of the hidden units, weights or biases are comfortably viewed as a *symbol*' (ibid., p.322).

So, these models do not have symbolic elements which group into structures with propositional content. There is no unique identification of a representation with a propositional content because in such a network 'there is no distinct state or part of the network which serves to represent any particular proposition.' (ibid., p.326). This becomes clearer when we consider that when information is processed by the network,

> '*many* connection strengths, *many* biases and *many* hidden units play a role in the computation. And any particular weight or unit or bias will help to encode information about *many* different propositions'.[3]

[2] As we saw in our account of the LOT hypothesis, Fodor alleged that propositions are expressed as LOT formulae in which symbolic elements are combined according to the rules of generative grammars. The view that representations are 'symbol structures' has remained one of the key commitments of the Classical Computational approach to cognition undermined by the model favoured by Ramsey et al. See Fodor and Pylyshyn, 1988, p.98 and Fodor and McLaughlin, 1990, p.202 for obvious evidence of the commitment and see the brief account of the systematicity and productivity debate below for an aspect of the rationale behind it.

That a representation of a state of affairs will have a unique identification with that proposition's content is also a commitment of Classical Computationalism. Fodor and Pylyshyn tell us that 'conventional architecture requires that there be distinct symbolic expressions for each state of affairs that it can represent' (Fodor and Pylyshyn, 1988, p.139) presumably because of the need for a single LOT formula at which analysis of a proposition will terminate if ambiguity is to be avoided.

[3] This facet of encoding in Connectionist models is often termed *superpositional storage*. As Andy Clark explains,

'Two representations are fully superposed if the resources used to represent item 1 are coextensive with those used to represent item 2. Thus, if a network learns to represent item 1 by developing a particular pattern of weights, it will be said to have superposed its representations of items 1 and 2 if it then goes on to encode the information about item 2 by amending the set of original weightings in a way which preserves the functionality (some desired input-output pattern) required to represent item 1 while simultaneously exhibiting the functionality required to represent item 2.' (Clark, 1993, p.17).

It is difficult to see how there can be fully superpositional *representation*. If the same set of units, weights, and biases can encode any number of propositions then the purported representations of propositional contents will be indistinguishable both structurally and functionally. How, then, can a superpositional representation be individuated? The reply that pictorial representations may be similarly indistinguishable (the duck/rabbit sketch, for example) is invalidated somewhat by the

Therefore, 'it simply makes no sense to ask whether or not the representation of a particular proposition plays a causal role in the network's computation' and, Ramsey et al maintain, 'it is just in this respect that our connectionist model of memory seems radically incongruent with the propositional modularity of common sense psychology.' (ibid., pp.326&327).

The departure from Classical Computationalism, with its concomitant propositional attitude Realism, is marked. The Connectionist model of memory presented by Ramsey et al appears to be devoid of any representational structure corresponding to a proposition. There are no elements, with the specifiable semantic content one might expect of a symbol, to form the structure. This is because the only candidates for these—the units, weights, and biases—can participate in the encoding of any number of propositions which bear no semantic relationship to one another. The lack of a representational structure means that there cannot be functionally discrete, semantically interpretable representations with a causal role within the cognitive system.

Ramsey et al demonstrate that there are no functionally discrete representational structures in their network (Network A) by explaining that were an identical network (Network B) to be trained to encode seventeen rather than sixteen propositions, the differences between the two networks would not be equivalent to a discernible structure responsible for the encoding of the extra proposition. As they put it 'these differences do not correlate in any *systematic* way with the functionally discrete, semantically interpretable states posited by folk psychology and by more traditional cognitive models' and the reason for this is that 'with information regarding any given proposition scattered throughout the network, the system lacks functionally distinct, identifiable sub-structures that are semantically interpretable as representations of individual propositions.' (ibid., p.328, my italics).

consideration that pictorial, but not superpositional, representations can be distinguished by virtue of their resemblance to what they depict. I mention this to highlight the problematic use of the word 'representation' in cognitive theorising, on which I will say more in the next chapter.

We can now see why, on the assumption that one of the original sixteen propositions scattered throughout the network is 'Dogs have fur', Ramsey et al tell us that, 'common sense psychology treats the class of people who believe that dogs have fur as a psychologically natural kind; connectionist psychology does not'. For if networks A and B are non-systematically different yet encode the same sixteen propositions, including 'Dogs have fur', then it follows that 'there are *indefinitely* many connectionist networks that represent the information that dogs have fur just as well as Network A does' and, because the differences are non-systematic, 'these networks have no projectable features in common that are describable in the language of connectionist theory'. So, 'from the point of view of the connectionist model builder, the class of networks that might model a cognitive agent who believes that dogs have fur is not a genuine kind at all but simply a chaotically disjunctive set' (ibid., p.329). Such a view is, of course, to be contradistinguished from the position Fodor has consistently occupied which insists that there are psychological natural kinds which can enter into nomic relationships (see the discussion of Fodor's antireductionism in the first chapter).

If this Connectionist model of memory and belief is correct, then there are no psychological natural kinds picked out by predicates of the form 'believes that *p*' or 'remembers that *p*'. Cognitive science should then reject propositional attitude Realism and give up its search for nomological generalisations 'couched in terms of the *semantic* properties of the attitudes'. Since Ramsey et al believe common sense psychology is a theory committed to propositional modularity and, hence, Realism, and their Connectionist model contravenes this commitment, their claim, that the correctness of the model entails the replacement of the common sense theory, would seem to follow. As we have intimated, should common sense psychology fall then Classical Computationalism, with its propositional attitude psychology and Representational Theory of Mind, would fall with it.

b) Connectionist Beliefs

Smolensky has resisted the Eliminativist argument not because he wishes to save Classical Computationalism but because he sees a need to talk of beliefs in explaining the activity in a network, such as that modelled by Ramsey et al, when it gives the output interpreted as 'true' on receiving a proposition as input (Smolensky, 1995, p.363). The notion of belief Smolensky develops is peculiar since he distinguishes two types of belief, to be found in a Connectionist network, without correlating his distinction with any made in common sense psychology.

Briefly, the Connectionist category of belief is divided into C-beliefs, specified by weight analysis, and L-beliefs, specified by learning analysis. A token of the former is a region of state space, and a network holds a C-belief when the weight vector encoding its knowledge lies within that region (ibid., p.369).[4] An L-belief is a single weight vector, encoding a proposition, within the region of the C-belief and it can be individuated only when the rest of the training set of propositions is specified (because it is specified as perpendicular to all the others) (ibid., pp.370-372).

The basic idea, then, seems to be that whether a network represents a proposition as true (and, therefore, can be said to believe that proposition) depends upon the numerical values encoding a proposition at the activation units, the value of the connection weights, and the contribution made by both to the value associated with the output unit. Beliefs, in themselves, are points or areas in an abstract space used to represent the aforementioned values. The

[4] As the technical details of Smolensky's notion of connectionist beliefs are of limited relevance to our concerns I shall not attempt a precise definition of state space. It will suffice to say that it is a *n*-dimensional space in which connection weights and input activity values provide coordinates for a weight vector **w** and an activity vector **a**.

If **w** is also represented as an arrow pointing to the weight vector from a point of origin, then a plane perpendicular to **w** will bisect the state space into a positive and negative half-space. The same operation can be performed to produce a plane perpendicular to an activity vector encoding a given proposition **p**. Thus, **p** is judged true if it lies within the positive half-space of **w** or when **w** lies within the positive half-space of **p**. The intersection of all the half-spaces for a training set of propositions is the solution space.

question Smolensky needs to, and does, address is whether these Connectionist beliefs are propositionally modular.

In order to understand Smolensky's answer to the question we need to appreciate that cognitive scientists and researchers into Artificial Intelligence generally accept that operations of computational/cognitive systems can be described at three levels; the *computational* level, the *algorithmic* level, and the *implementational* level.[5] At the computational level of description we are given the function that the system computes (which will, in the case of human cognition, tend to be expressed in the intentional idiom, that is, at the semantic level). At the algorithmic level the means by which the function is computed are given, while at the level of implementation the description tells how the function is physically realised, that is, in what medium and by which physical processes in that medium.

Smolensky's suggestion is that at the computational level of description, of the network proposed by Ramsey et al, we can give Connectionist beliefs a semantic interpretation and show them to be functionally discrete. As far as semantics are concerned, the functions of the network are defined in terms of weight and activity vectors which are semantically interpretable (because a given activity vector can be said to encode a given proposition) and since Smolensky's C and L-beliefs are also defined in these terms it would follow that they are also semantically interpretable (ibid., pp.374-377). The story regarding functional discreteness is a little more complex.

Two networks, A and B, will have weight vectors - **wA** and **wB** respectively - which lie within the solution spaces - **SA** and **SB** - (see note 4 above) for the propositions encoded by each network. Network A is to be that of Ramsey et al and will encode sixteen propositions, while Network B is a similar network encoding the same sixteen and an additional proposition. **SB** will be smaller than **SA** (since the former encodes seventeen and the latter sixteen

[5] The levels were distinguished by David Marr in *Vision* (1982) where he develops a theory of visual perception.

propositions) so, remembering that for a network to have a C-belief is for its weight vector to lie within the positive half-space produced by the encoding of the relevant proposition,

> 'if some process were to disturb the weight vector **wB**, so that it moved out of the solution space **SB** while still remaining within the larger solution space **SA**, it would make perfectly good sense to say that the second net had "lost" or "forgotten" the 17th belief, while retaining the other 16.' (ibid., p.375).

That a belief can be added or lost without disturbing the rest of the belief store was, of course, a criterion for judging it to be functionally discrete (see 1) above). Furthermore, Smolensky points out that the C-beliefs of A and B are projectible predicates (that is, predicates which can appear in nomic statements) describing the knowledge-encoding vectors **wA** and **wB** (ibid., p.374).

Thus, at the computational level we can specify C-beliefs as both semantically interpretable and functionally discrete. (Though L-beliefs have the former property, there is some doubt as to whether they have the latter (see ibid., p.375, and Macdonald, C., 1995, p.302).) However, since Connectionist beliefs are not physically located in spatial sub-regions of the networks (Smolensky, 1995, p.374), though they are functionally discrete at the computational level they are not at the algorithmical, nor the implementational, level but it is at these levels that we describe the causal processes of the network. Consequently, we cannot assign a causal role to C or L-beliefs, for there is nothing functionally discrete at the algorithmic level which can be called a belief and, therefore, nothing at the implementational level with the causal role of a belief. This would put Smolensky's thesis into that area of propositional attitude Realism inhabited by Epiphenomenalism—an area most philosophers of mind would avoid.

Smolensky's ingenious abstraction of beliefs, from Connectionist models, would appear to have gained little ground on the Eliminativism of Ramsey et al since, by his own admission, it falls short of the requirements of propositional modularity, the fulfilment of which license the attribution of propositional attitudes to a system. However, as he sees it, the notion of belief is required in order for Connectionist theory to explain a network's ability to judge propositions. He says of Ramsey et al,

> 'What they have is simply a network. When we turn to actual Connectionist theory to explain this network's behavior, to see what explanatory notions are thereby made available to Connectionist psychology, we see that in fact these notions *do* redeem Classical belief - partly.' (ibid., pp.378&379).

The retreat indicated by the word 'partly' is due to the inability to individuate a belief at the algorithmical level of description, which would be possible in Classical models.

So Smolensky seems to have accepted the claim of Ramsey et al that folk psychology is committed to propositional modularity, and he admits that Connectionist networks like theirs will realize only two of the three properties constitutive of propositional modularity, but he is not willing to accept the consequent of the conditional claim, *viz.*, that Eliminativism is correct.

In 'Connectionist Minds' (Clark, 1990) Andrew Clark, like Smolensky, suggests that the Eliminativist argument of Ramsey et al gains plausibility by restricting description of the activity of Connectionist networks to the level of units and weights. Again like Smolensky he maintains that the folk psychological notion of belief is applicable to such networks at a higher level of description. However, whereas Smolensky appeals to analyses of activity and weight vectors as a means of individuating belief states, Clark argues that *post hoc* statistical analysis of certain types of network yields cluster profiles of its activity which are not only semantically interpretable, but also distinguish the activity which causes the networks output (see ibid., pp.345-348). A state which is semantically interpretable and discrete in its causal efficacy would seem to be apt for description by a projectible psychological predicate, that is, a predicate amenable to the nomic generalizations of folk psychology (see 2) above). The networks Clark considers (which are more extensive than those discussed by Ramsey et al) will be conducive to propositional modularity and, therefore, lend credence to folk psychological theory.

Thus we see that, whilst Smolensky accepts that Connectionist beliefs lack propositional modularity but rejects the Eliminativist consequences, Clark argues for propositional modularity of Connectionist beliefs—at least, this is the way Stich and Warfield characterise the positions in a reply to Smolensky and Clark (Stich and Warfield, 1995, p.395). Without going into detail, the two main difficulties with Clark's account of Connectionist beliefs, according to Stich and Warfield, are firstly, that he over generalizes—the cluster profiles apply only to a restricted class of networks (ibid., pp.398&399)—and, secondly, he seems in danger of characterising belief as a behavioural phenomenon, and this should be anathema to anyone attempting to explain cognition as a computational process in the central nervous system (ibid., pp.402&403).

Regarding Smolensky's rejection of the Eliminativist conclusion of the argument offered by Ramsey, Stich, and Garon we find, surprisingly, that Stich and Warfield are in agreement albeit for rather different reasons. Their reasoning is that 'there is a significant logical gap between the claim that folk psychology is seriously mistaken, and the claim that the propositional attitudes to which folk psychology appeals do not exist' (ibid., p.405), and that the premisses which might be assumed in order to fill the gap are 'deeply problematical'.

The first of the two premisses they consider is the description theory of reference. In this context the theory explains that terms of the folk psychological theory refer to propositional attitudes via a class of descriptions entailed by the theory. The explanation would facilitate the Eliminativist conclusion because if the folk theory was seriously mistaken, it would entail the elimination of the propositional attitudes. This is because the descriptions will be demonstrated to be false and the terms of the theory, thereby, will be deprived of their reference. A version of the description theory is attributed to David Lewis (for whom the descriptions are of the causal roles of the referents of the theoretical terms (see Lewis, 1972)) but it clearly has a Russellian provenance.

Stich and Warfield point out that the description theory has fallen into disrepute largely as a result of arguments and examples advanced by Putnam and Kripke who have proffered, what may be called, a causal/historical account of reference.[6] On this account the reference of a term is fixed by an initial 'baptism' of a referent (or class of referents if it is a natural kind term) such that on subsequent correct use of the term its reference is determined by a causal chain connecting current users to the baptism. Consequently, though folk psychology may be radically mistaken it could not be concluded that there are no referents of its theoretical terms. Stich and Warfield point out that, although the ancients, who believed that stars were holes in the celestial dome have been proven wrong, it makes sense to suppose, on the causal/historical theory, that they were talking about the same 'heavenly bodies'—perhaps 'light sources' would be better—as modern astronomers (Stich and Warfield, 1995, p.407).[7]

Short of an adequate defence of the description theory and a refutation of the causal/historical theory, it would seem that Eliminativists will have to look elsewhere for a bridging premiss to fill the logical gap between folk psychology's falsity and their conclusion. The alternative offered by Stich and Warfield is to hold that propositional attitudes have conceptually necessary, or 'constitutive' properties (ibid.). Given the previous discussion, it might be profitable for the Eliminativist to insist that having propositional modularity is constitutive of a state's being a state of belief or desire. The failure to demonstrate that

[6] It is noteworthy that the arguments applied by Putnam and Kripke have had direct as well as indirect relevance (that is, via the rebuttal of the description theory of reference) to issues in the philosophy of mind. Acceptance of Kripke's view of natural kind terms as rigid designators (as opposed to descriptions) presents problems for Type Identity theorists, and Putnam's 'Twin Earth' argument has worried Internalists trying to account for mental content.

[7] Of course, the fact that it makes sense to suppose that both ancients and moderns can use terms with the same reference, while describing their referents in very different ways, would create a substantial problem for the description theory of reference.

Connectionist beliefs are propositionally modular would then supply the logical link between the correctness of Connectionist models and Eliminativism.

Having offered the premiss Stich and Warfield proceed to discredit it. Of the two reasons for questioning its validity the first is that the decision as to what properties are constitutive of a natural kind is somewhat arbitrary (ibid., p.408). The second is that the supposition that some properties are conceptually necessary seems to rely on a distinction between analytic and synthetic sentences since the sentence 'beliefs and desires are propositionally modular' would have to state an analytic truth. However, since Quine is considered, by many philosophers, to have successfully denied that there is a hard and fast distinction between analytic and synthetic sentences, the Eliminativists must refute the arguments for this denial if they are to be allowed their conclusion. Though they would not preclude such a refutation Stich and Warfield are doubtful it will be imminent (ibid., p.409).

The outcome of the debate on the relationship between Connectionism and Eliminativism, for Stich and Warfield, is that 'connectionism *might* show that commonsense psychology is wrong, but it lends no support whatever to the claim that common-sense psychological states do not exist' (ibid., p.410). For Stich this constitutes a retrenchment not only on his claims for Connectionism but also on his ontologically radical conclusion based on the experimental evidence which suggested to him that '*there are no such things as beliefs*'. Indeed, in 1983 Stich was amongst those who reached Eliminativism via a description theory of reference (see the last section of the previous chapter). That said, his more moderate proposal of excluding the folk psychological categories from cognitive theorising, as a result of the Infralinguistic and 'Wittgensteinian' considerations, would not be undermined by his 1995 thesis. The moderate proposal does, however, share with Eliminativism the paradoxical position of advocating a science of cognition devoid of cognitive concepts.

One way of characterising Stich's recent position would be as a midpoint between propositional attitude Realism and Eliminativism. His assumption that there are no propositionally modular beliefs is tantamount to a denial of Realism, yet he does not allow that this permits the further step to Eliminativism. This *appears* to be a paradoxical position given that Realism involves a commitment to the existence of mental states and its denial would seem to entail their elimination from the class of existent entities.[8] Admittedly there is a Realist position which would permit propositional attitudes to lack causal efficacy (in the vein of Smolensky's Connectionist beliefs), and therefore lack propositional modularity, but this is Epiphenomenalism and it is unlikely Stich wants to be associated with this.

Even if the reasoning of Stich and Warfield is right and the correctness of certain Connectionist models does not entail Eliminativism, the lesser claim, that it does entail the falsity of folk psychology, will still have an impact within cognitive science. It bodes ill for Classical Computationalism which relies on folk psychology as the source of its nomic generalizations. The falsity of the folk theory would make this reliance seem ill-judged. However, it might be that a characterization of Connectionist beliefs, which meets the criteria for propositional modularity, will be forthcoming, but even in this eventuality, and even if the characterization were accepted by most cognitive theorists, there would still be a tension between the Classicist and the Connectionist. The reason for this is the disparity between the two types of representational system postulated by each approach. I will now give a brief account of the Classicist/Connectionist debate.

[8] Note that the paradox is removed if one rejects the view that psychological terms are used to refer to internal states with causal and/or semantic properties. Unfortunately Stich, in his insistence that folk psychology is a theory, cannot reject the view.

2 CONNECTIONISM AND CLASSICAL COMPUTATIONALISM

a) The Compositonal Thesis

The debate between Classicists and Connectionists revolves around the question of whether Connectionist architectures support the postulation *of composite representations made up of context-independent elements*. The Classicists say that if they do then Connectionist models are merely implementations of Classical ones, and if they do not then Connectionist models are lacking three key features of cognition, namely; systematicity, inferential coherence, and productivity.[9]

Once again, let us cast Fodor as our typical Classicist. As we have acknowledged, Fodor has it that a propositional attitude is a computational relation to a representation expressing the relevant proposition. The representation has a constituent structure which consists of elements corresponding to the elements of the proposition, and herein lies the representational relationship. Fodor and Brian McLaughlin express this as follows;

> 'If a proposition *P* can be expressed in a system of mental representation *M*, then *M* contains some complex mental representation (a "mental sentence") *S*, such that *S* expresses *P* and the (Classical) constituents of *S* express (or refer) to the elements of *P*.' (Fodor and McLaughlin, 1990, p.202).

A symbolic element is a Classical constituent of a representation only if the former is tokened whenever the latter is. The example provided is that of the 'mentalese' symbol which names John which is necessarily contained in the mentalese symbol that means that John loves the girl (ibid., p.201).

That Classical constituents are context-independent is demonstrated by the fact that the same constituent can express the same content in the context of any sentential representation in which it occurs. Regarding the above example, tokens of the representation type *John* will

[9] Fodor has argued elsewhere that the advocate of Intentional Realism must also accept the language of thought hypothesis because it is only the two combined that will account for these features of cognition (see Fodor, 1987 appendix).

have the same content in representations of the sentence 'John loves the girl' and 'The girl loves John' (ibid., p.207).

As we noted in our earlier exploration of Fodor's LOT hypothesis there is an advantage to be gained from postulating mental representations composed of elements because when these are combined with an innate generative grammar we can provide an account of the ability to produce and understand an unbounded set of natural language sentences without the need for a distinct representation corresponding to each member of the set. In other words, the Classical thesis that representations are composite explains the *productivity* of cognition (see Fodor and Pylyshyn, 1988, p.116).

Fodor and Pylyshyn argue that the *systematicity* of cognition provides another motivation for the Classical commitment to a combinatorial syntax and semantics of representations. They tell us,

> 'what we mean when we say that linguistic capacities are *systematic* is that the ability to produce/understand some sentences is *intrinsically* connected to the ability to produce/understand certain others' (ibid., p.120).

Thus 'no speaker understands the form of words "John loves the girl" except as he/she also understands the form of words "the girl loves John"', and what goes for linguistic capacities goes for thought also because 'you don't find people who can *think the thought* that John loves the girl but can't think the thought that the girl loves John' (ibid., p.121). This feature of language and cognition seems readily explicable if representations expressing propositions are complexes of context-independent atoms able to occur in any number of complexes, depending upon the proposition requiring expression.

A further advantage of Classical architectures is that they account for the fact that a person who understands a compound proposition of the form *p&q&r* can infer that *p* and can also infer *p* from *p&q*; that is, they account for *inferential coherence*. This is possible in a Classical model in which the representations of propositions are composite because 'the process of drawing the inference corresponds, in both cases, to the same formal operation of detaching the constituent that expresses the conclusion' (ibid., p.129). Expressed in this way it is clear that inferential coherence and systematicity both require of any system that can express a proposition that it can also express the proposition's elements.

The argument of Fodor and Pylyshyn is that these three features of genuine cognitive systems are found in Classical models but not Connectionist ones in which representations are either 'labelled' nodes (units) or are distributed over such nodes. The first type of representation lack the requisite constituency because the nodes are always, in fact, atomic (even if their labels are not) and therefore lack constituents. The second type—the distributed structures—are usually composed of micro-feature vectors. A micro-feature is realized by an activity vector which might, for example, express the predicate 'is a hot liquid' when the distributed representation is of the concept **coffee** (see Smolensky, 1991, pp.172&173). However, such a micro-feature is not a constituent in the way that the symbol representing John is in the mentalese sentence 'John loves the girl' because 'is a hot liquid' has only a definitional relationship to **coffee** and 'it really is very important not to confuse the semantic distinction between primitive expressions and defined expressions with the syntactic distinction between atomic symbols and complex symbols' (Fodor and Pylyshyn, 1988, p.105). After all, 'the definition relation can hold in a language where *all* the symbols are syntactically atomic'. In addition, the actual value of the activity vector will vary depending on whether it contributes to the representation of **cup with coffee** or **man with coffee**. For this reason the micro-feature is context-dependent rather than independent (see Smolensky, 1988, pp.67&68) which would imply that the ability to understand **cup with coffee** is not intrinsically connected to the ability to understand **man with coffee** in Connectionist systems.

Since Connectionist representations do not have the requisite combinatorial syntax and semantics there cannot be Connectionist systems which display productivity, systematicity, and inferential coherence in their processing. Moreover, if there were Connectionist representations of the required sort they would have to differ, in some marked way, from Classical representations, for otherwise it would seem that the most Connectionism could provide is a new lower-level description of how Classical models can be implemented. This is considered unsatisfactory for those Connectionists who see their models as theories of cognition which improve on, or refine, the Classical models (see Fodor and Pylyshyn, 1988, p.149&150).

b) Connectionist Representations

The challenge Smolensky accepts is to provide an account of how Connectionist representations are composite and have context-independent constituents. Again, comprehensive coverage of Smolensky's theorising would exceed our requirements so I will present merely the penumbra of Smolensky's account of Tensor Product Representations.

Given that the supposed strength of Classical architectures is that they are able to model mental processes sensitive to the compositional structure of the representations over which they range, the problem for Smolensky is to show how complex structures, such as parse trees, can be mapped onto vectors of activity in Connectionist networks. Smolensky tells us that,

> ‘a general formal framework for stating this problem is to assume that there is a set of discrete structures S (like parse trees) and a vector space V - a space of activity states of a Connectionist network’.

‘A Connectionist representation’ then ‘is a mapping from S to V’ (Smolensky, 1991, p.178). The favoured candidates for the mapping are tensor product representations which facilitate the mapping. We are told that the tensor product representation is constructed as follows:

Firstly, the discrete structure is decomposed into constituents, each filling a particular role within the structure; secondly, two representations are specified—one for structural roles, and the other for the constituents (or fillers). The role-sensitive representation of a constituent part of the structure is built by taking a particular vector product (of the vector that represents the constituent *independently* of its role) and performing the tensor product operation (vector multiplication) on this and the vector representing the role in the structure filled by the constituent (ibid., p.179). The representation of the whole structure is built from the representation of its constituent parts by the operation of superposition (vector addition). (For a simple explanation of the superposition and tensor product operations see Smolensky, 1995a, p.238).

So by availing him or herself of the tensor product operation the Connectionist is able to combine vectors representing the role of a constituent and the 'filler' of that role to produce a representation of a constituent proper. The composite structure can be represented by a vector produced by the superposition operation when carried out on the tensor product vectors representing the constituents. A 'filled role' vector is a tensor product which might represent the word 'John' by combining activity vectors representing *roles* for letters, such as *the first letter position*, and activity vectors representing the letters themselves, in this case 'J'. Thus a tensor product will represent *'J' in the first position*. Since the tensor product operation is recursive (it can operate on tensor product vectors) the tensor product representing a word can be composed of those representing letters (see Fodor and McLaughlin, 1990, p.211&212).

So, is it the case that Smolensky has succeeded in providing Connectionist representations which are combinations of context-independent constituents and which satisfy the need for systematicity in a cognitive system?[10] Fodor and McLaughlin think not. However, it is beyond

[10] The debate concentrates on the systematicity requirement as the need for productivity is questionable, and the need of inferential coherence does not place any tighter constraints on the cognitive theorist than that of systematicity (see Fodor and Pylyshyn, 1988, p.116ff).

the needs of this exposition to delve into the intricacies of this debate. I will conclude it by mentioning that Smolensky's reason for arguing that tensor product representations are not simply implementations of Classical constituents turns out to be one of the reasons Fodor and McLaughlin cite for his failure to solve the Connectionist's problems with systematicity.

In the case of Classical constituents, each will have both a causal (syntactic) and a semantic role simultaneously because the formal entities constitutive of the representations will be semantically interpretable whilst also supporting the cognitive processing involving those representations. With Smolensky's Connectionist constituents, however, this is not the case because 'formal, algorithmical specification of processing mechanisms, on the one hand, and semantic interpretation on the other, must be done on different levels' (Smolensky, 1991, p.167). In other words, the representation of a propositional content is specified at the higher level of vector analysis while the processes producing the system's output are fully specified only at the, lower, algorithmical level defining the interactions of the individual units and their numerical values. Indeed it is the individual activity values, and not the vectors which they comprise, 'that really drive the machine' (ibid., p.190). Thus, Smolensky's representational constituents are not causally efficacious because the level at which they are identified is not the level of causal processing. It is this fact that is taken by Smolensky to warrant the assertion that Connectionist models are not mere implementations of Classical ones (ibid., p.167).

Anticipating objections, he claims that the causal inefficacy of his constituent representations cast no more doubt on their credibility than it does on representations of states of constituent electrons in atoms (as tensor products of 'spin' and orbital role vectors) where operations on vectors are not what *cause* the changes of the atom's states (see ibid., p.196).

Fodor and McLaughlin, however, cannot see how causally inert constituents can account for systematicity as 'the explanation of systematicity turns on the causal role of the constituents of mental representations' (Fodor and McLaughlin, 1990, p.218). Assuming that it is necessarily the case that if a machine can represent the constituents a, R, and b it can also represent the propositions aRb and bRa, (that is, given the systematicity requirement) what is needed is an explanation of how the machine guarantees the nomic relation between the representation of the constituents and the production of the two propositions.[11] This just is the problem of systematicity and Fodor and McLaughlin cannot find an answer to it in Smolensky's brand of Connectionism (see ibid., p.215). However, they are content that Classical Computationalism provides the solution by giving the representational constituents a causal role as well as a semantic content. It is because of this role that 'mental processes are sensitive to the constituent structure of mental representations' (ibid., p.203) and it is this sensitivity which affords the necessary systematicity. Another way of expressing this might be to say that the aforementioned nomic relation is guaranteed by the machine's manipulation of symbols determining their own causal role.

An additional consequence of the fact that processing in Smolensky's model is carried out at the sub-symbolic level is that the laws generalising over the processing will describe phenomena which are not semantically interpretable. Obviously propositional attitude predicates, which describe (as Fodor has it) relations to semantically interpretable representations, will be excluded from these laws and at best relegated to imprecise, and hence non-lawlike, generalisations (see Smolensky, 1988, p.42 for his insistence on the imprecision of symbolic level descriptions of sub-symbolic processing—a case of hoist by one's own petard). The result is that, once again, Smolensky has led Connectionist theorising into the realms of Epiphenomenalism (in this case Content Epiphenomenalism), for the fact that Jack understands the sentence 'John loves the girl' cannot play any part in the ætiology of his

[11] Fodor and McLaughlin say that the crux of the problem faced by Smolensky is 'to explain how sytematicity could be *necessary*—how it could be a *law* that cognitive capacities are systematic' (ibid., p.216). Hence my use of the words 'necessary' and 'nomic'.

thought 'So the girl is loved by at least one person' and will not licence the prediction that he will also understand the sentence 'The girl loves John'.

3 BETWEEN ELIMINATIVISM AND REALISM

The debates we have examined suggest that Connectionists must accept one of two alternatives. Either they must aim to produce models of cognition which do not employ semantically interpretable, functionally discrete, causally efficacious representations, in which case they will not be modelling propositional attitude states and will not, therefore, be propositional attitude Realists. Or they can claim that their models use representations and have propositional attitudes but concede that these have no role in the ætiology of behaviour or even other mental states and events. The result is Epiphenomenalism which, in denying that mental phenomena or mental properties have causal powers, puts cognition outside the realms of nomological generalization (Psycho-Physical Parallelists notwithstanding) and this is, one would have thought, a dismal prospect for a science of cognition.

However, there are those who align themselves with the Connectionist project who would deny it is condemned to either of these alternatives. They might characterise theirs as a position between Realism and Eliminativism and what this will amount to is a belief that folk psychology has a role in explanations of cognition but that it does not employ a taxonomy the terms of which refer to anything corresponding to an internal state of the cognising subject. We need look no further than Clark (1993) for advocacy of such a position.

After recounting Stich's argument against the folk psychological taxonomy based on evidence for there being separate verbal and non-verbal cognitive storage systems (see Stich, 1983, pp.230-237) Clark writes,

> 'In the inner realm, folk-psychological items fragment. Things fall apart. The combined fluidity and fractionalbility of the inner economy (relative to the folk

> ontology) is nothing short of astonishing. But it can be astonishing, confounding all our natural expectations, without compromising the folks' explanations of actions or the ontology of folk solids.' (Clark, 1993, p.201).

To understand this claim we need to look briefly at what Clark says about the 'folk solids' and explanations of action.

Clark uses the expression 'folk solids' to refer to concepts, propositions, and attitudes (ibid., p.3) and it is in reference to concepts that he begins to limn the intermediary position we are considering. His suggestion (which he admits is not original) is that to ascribe grasp of a concept to someone is to ascribe a general skill 'which may (consistent with the validity of the folk ontology and explanations) depend on one of a variety of computationally disunified subskills' (ibid., p.203). He adds that,

> 'by "disunified" I mean that the subskills need display no unity visible without the lens of folk-psychological interests. Instead they amount to a bag of tools, some of which may be verbal, some imagistic, some sensorimotor, and so on.' (ibid.).

So ascribing a concept is like ascribing a skill (such as the ability to play golf) and, consequently, 'the requirement of inner scientific unity is misplaced for the simple reason that ascribing a concept does not commit us to the presence of any associated unitary and recurring inner state' (ibid., p.204). In a similar way the ascription of an ability to play golf, which involves a variety of accomplishments (such as those of putting, driving, chipping, and so on), need not be based on the assumption that there is some common neurophysiological system responsible for all these accomplishments and, therefore, constitutive of a person's ability to play golf. Just as Realism regarding the ability to play golf seems untenable so, Clark would have it, Realism regarding the ability to use a concept should also.

Clark suggests that in ascribing the grasp of a concept to a person,

> 'we are really ascribing a body of knowledge and skills whose manifestations may be both internally disparate (in terms of occurrent representational states) and externally

> disparate (in terms of, e.g., abilities and verbal and nonverbal skills).' (ibid., pp.204&205).

I think the idea is that the behaviour (or subskill) which manifests the grasp of a concept (such as applying the concept appropriately in conversation and responding appropriately to its use) is produced by an internal process (which might also be a subskill in Clark's terminology) involving occurrent representational states. These behaviours and processes, though disparate, are identified as constitutive of possession of a concept by a folk psychology for which they emerge 'against the biological and cultural background of human needs and institutions' (ibid., p.205).

Insofar as Clark is saying that ascription of a concept to someone usually depends on his or her manifest ability to use a word and to respond appropriately to its use, I am in agreement. However, I disagree with his apparent presumption that this is an insufficient explanation of concept possession and that we need to postulate knowledge stores and internal processes ranging over occurrent representational states if we are to approach a sufficient one. Presumptions of this sort are the *raison d'être* of cognitive science which generally tries to explain intelligent behaviour by recourse to inner (ultimately neural) processes. I will say more about this in chapter 7.

The foregoing provides an account of Clark's rejection of Realism regarding concept ascription. Regarding the other 'folk solids', namely propositional attitudes, he adheres to Ascriptivism according to which, roughly, beliefs, and the like, are attributed on the basis of observable patterns in actual and counterfactual behaviour. However, though he thereby eschews Realism about *beliefs* he advocates Realism about *believers* (ibid., p.216). He has two reasons for adopting such a potentially paradoxical position and both are constraints on the '*class of beings* for whom mentalistic interpretation is appropriate'. One of these is the 'normativity requirement' by which folk ascriptions must be defeasible.[12] In other words, we need a Realist construal of the class of believers to explain 'our ability to judge our own performance as either living up to our antecedent commitments or failing to live up to them' or put simply, we need to explain how we 'judge our judgings as correct or mistaken' (ibid., p.217).

To see why Clark advocates Realism about believers we might contrast his view with that of Dennett who opts for Ascriptivism about believers.[13] According to Clark, the trouble with Dennett's position is that it is vulnerable to counterexamples such as that of the Giant Lookup Table,

> 'a being whose actual and counterfactual behavior is exquisitely honed to display the patterns characteristic of grasping concepts and acting on the basis of beliefs but whose inner computational life (consisting of an astoundingly large collection of preset responses to specific inputs and input sequences) seems curiously inappropriate for a True Believer' (ibid., p.214).

The Giant Lookup Table may behave like a true believer even in its admission of an error in judgement but, according to Clark, it is not a true believer because such an admission 'is in no way traceable to any internal process which retrieves a memory of the previous judgement and assesses it against the backdrop of the system's general knowledge and commitments' (ibid., p.217). So, what lies behind our intuition that the Giant Lookup Table does not correct its judgements is 'the total lack of any internal mechanism by which traces of earlier outputs can

[12] The other requirement is that of consciousness because 'it seems...conceptually impossible for a being to count as grasping a concept and yet be incapable of ever having any conscious experience involving it'. A complete cognitive science, then, will incorporate a theory telling us what makes conscious qualitative experience possible (Clark, 1993, pp.217&218).

[13] Ascriptivism is part of Dennett's characterisation of the Intentional Stance which, he maintains, we adopt when we view physical systems as having propositional attitudes. An Intentional System is one whose behaviour can be predicted and explained by ascribing propositional attitudes (see Dennett, 1971) although it does not follow that these correspond to any internal states. This allows Dennett to refrain from adherence to the strong Realism we have been dealing with (see Dennett, 1987a, p.69).

At the same time, according to Dennett, adoption of the Intentional Stance is practically unavoidable with regard to oneself and one's fellow intelligent beings (Dennett, 1981, pp.25-27) and consequently it is not possible to eliminate folk psychology (Dennett, 1987b, pp.233-235). For these reasons it would be correct to say that both Dennett and Clark opt for a position between Realism and Eliminativism. Both believe a scientific explanation of cognition is required and that the postulation of mental representations will be needed for this. Furthermore, Dennett also seems optimistic about the potential of Connectionist theorising to provide the requisite explanations (as we noted during our glance at the 'frame problem' in section 2 c) of the previous chapter).

be stored and later reassessed'. It is interesting to note that, for Clark, the normativity requirement demands an internal mechanism. The incompatibility between the normative and the mechanical is a theme I shall explore in chapter 6. Clark rejects Classical Computationalism's Realism, and hence its concomitant LOT hypothesis, but he suggests that Connectionism is a version of the Representational Theory of Mind (ibid., p.235, note 1). Demonstrating the incoherence of that theory will be the task of the next two chapters.

So, although Clark has disavowed propositional attitude Realism he is committed to the existence of mechanisms for the storage and retrieval of knowledge and this appears to be the rationale behind his advising Dennett to be a 'realist about believers' (ibid., p.216). To complete the characterization of Clark's position as lying between Realism and Eliminativism we should highlight a difference between Clark and the Eliminativists.

Whilst Clark, in denying Realism, has denied that beliefs are propositionally modular and has, therefore, implied that they have no causal role, he *is* freed from the charge of Epiphenomenalism, levelled at Smolensky, because Epiphenomenalists are Realists—they assert the existence of mental phenomena. However, should he attempt to vindicate folk psychology, he might be accused of inconsistency. This is because, as the Eliminativists see it, folk generalisations pose as causal explanations ranging over mental states and processes, and one cannot hold this view whilst also accepting that there is a lack of any 'inner items which *realize* the propositionally described states and do the causing' (ibid., p.210). In other words, if one is not a Realist then one must admit that the folk theory is false and apt for elimination.

Clark suggests that the correct response to the Eliminativist's demand for an account of propositional modularity 'is to reject outright the idea that folk psychology is necessarily committed to beliefs and desires as being straightforwardly causally potent' (ibid., p.211). Talk of beliefs and desires may have explanatory value even though there is no 'specific underlying scientific essence' with which to identify them. Clark's suggestion is that we can get psychological predicates to give non-straightforwardly causal explanations of an event using counterfactuals and, what amounts to, the method of differences. For example, when I buy a drink on a hot day we can say it was an affect of my belief that the drink is cold rather than the belief that dogs have fur because if, in close possible worlds, I had the first but not the second belief I would still buy the beer but not vice versa (ibid., pp.211&212). Thus, highlighting a belief in an explanation of behaviour may be counterfactually justified.

Though he concedes the anticipated objection that 'the counterfactually warranted highlighting of an individual belief falls short of establishing it as a genuine cause' Clark responds by saying that,

> 'an additional argument, to the effect that all *good* explanation must be straightforwardly *causal* explanation, would be needed to amplify this concession into a fatal objection to the folk understanding of mind. Even a brief reflection on the varied panoply of human explanatory projects should convince us that no such general claim can be sustained' (ibid., p.213).

At an earlier point Clark maintains that rejection of Realism about 'folk solids' should not lead us to conclude that the common-sense ontology is incoherent or false, 'quite the contrary. The folk explanations simply occupy a different arena' (ibid., p.207). He does not, however, describe that arena.

Again, I find myself in agreement with Clark up to a point. I also do not believe that psychological explanation is 'straightforwardly' causal but for reasons different from those Clark seems to have. His apparent belief that such explanation has some causal quality because it can be counterfactual supporting is no doubt parasitic upon the fact that, on many modern characterizations of causation, counterfactuals are appealed to in defining causal relationships. Perhaps, by mimicking causal explanations, psychological sentences can acquire explanatory value by association, but the Eliminativist would have a good case for insisting on explanations which refer to genuine causes instead of psychological constructs. It would make more sense, as far as I can see, to follow the Rylean lead hinted at in the last quotation above and deny that psychology is in the business of causal explanation in the sense of *efficient* causation. Of course, to do so would entail rejecting the belief that there is a folk-psychological theory, a rejection that would be heresy to the Eliminativists (and many Computationalists also). By Rylean lights a mechanistic account of cognition would constitute a category mistake and, since Clark seems to be advocating such an account, it is not surprising that he struggles to create a whiff of causation in psychological explanation.

CHAPTER FOUR

REPRESENTATIONS AND THE INFINITE REGRESS ARGUMENT

The last three chapters have presented an overview of some of the main positions and debates arising within the philosophy of cognitive science. The task of the next four chapters is to reveal the conceptual distortions involved in viewing the science of cognition as a viable enterprise. This task is not best accomplished by dealing with the various positions individually for, although cognitive theorists group around certain basic assumptions, within these groupings there are differences in argumentative strategies which make it seem that the positions of the individual theorists, like snowflakes, are never identical. For this reason I will try to make the task more manageable, firstly, by tilting at the assumptions which seem to be common to most, if not all, cognitive theorists and, secondly, by taking aim at those champions of the assumptions who have been the most influential in their field.

The general form of my critique is to highlight problems with those assumptions and then select what I take to be the best, or the only, form of solution available from within the cognitivists' theoretical edifices. I then argue that what are available are not solutions but, at best, problems masquerading as solutions. By the end of the thesis I hope to have demonstrated that the assumptions a science of cognition *must* make cannot be made without distortion. This distortion involves using certain words and expressions in extraordinary contexts while, illicitly, importing into those contexts the connotations those words and expressions ordinarily have.

The first assumption, to be dealt with in the current chapter, is that we can speak intelligibly of internal *representations*. As I shall explain shortly, Realists within cognitive science need to posit internal representations in order to account for the content of propositional attitudes states. But the concept of 'representation' is pressed into service by Irrealists as well, whether they be Eliminativists or those, like Clark and Dennett, who wish to occupy a position between Realism and Eliminativism.[1] In the case of Realism the notion of internal representation is set within the framework of the Representational Theory of Mind and, as such, the notion is employed in outlining the nature of all mental phenomena with the property of Intentionality (also to be examined shortly). The Eliminativist conception of internal representation will not have a home in that framework, for the Eliminativist will not explain, but will try to explain away, mental phenomena. However, this more minimal conception of internal representation will be equally susceptible to the charge of vicious regress which will be pressed later in this chapter. It should be borne in mind that, since it denies the legitimacy of intentional concepts (such as 'understanding' and 'interpretation'), Eliminativism cannot conscribe them in order to keep its account of representation afloat. As we shall see, Realism cannot appeal to these concepts either, when justifying its representationalism, unless it can describe their signification in naturalistic terms—in terms, that is, which are not drawn from the intentional idiom.

Indeed, the assumption that we can naturalize intentional content is another of the main targets of the forthcoming critique. It is the subject matter of chapter 5 and, since I argue that this assumption presupposes another, *viz.* that normativity can be naturalized, it is indirectly the target of chapter 6, which argues that accounting for the normativity of language and cognition reintroduces the very concepts cognitive science purports to explain. Since the naturalization of normativity is prerequisite for the naturalization of intentional content (and is, therefore, also prerequisite for use of the concept of 'representation' in cognitive science), and the former naturalization cannot be accomplished in advance of the

[1] For an example of the Eliminativist use of 'representation' see P.S. Churchland and T.J.Sejnowski, 1989, p.247. Here Churchland and Sejnowski consider levels of neural representation having speculated on how such representations might be modelled by Connectionist networks.

latter, the programme of bringing cognition within the explanatory reach of a natural science begins to look highly suspect.

Unfortunately, the line of argument pursued is not quite as straightforward as the previous paragraph might suggest. This is because I try to cover those exits from the argument which I envisage being opened by its key targets. For this reason chapter 7 provides a more general survey of, what I take to be, the damage done to the idea of a cognitive science by the arguments of chapters 4,5, and 6 (although, even here, I find it necessary to engage, at closer quarters, with one or two theses). In the last chapter I return to the Eliminativist thesis which, since it does not explicitly require an account of internal representation or intentional content, is not a primary target for many of the earlier arguments. However, we do find that some of the considerations, regarding normativity, have direct relevance to both the conclusion and one of the premisses of the Eliminativist argument.

1 THE REPRESENTATIONAL THEORY OF MIND

As I suggested, the concept of 'representation' and, by association, the concepts of 'symbol', 'sentence', and 'formula', is an explanatory tool the cognitive scientist can rarely do without. It is most heavily used by the Realist who, in relying on it to fashion his account of intentional content, develops it into the Representational Theory of Mind.

Since there are a variety of versions I should like to make clearer what I take to be the Representational Theory of Mind (I will occassionally shorten this to RTM, or representationalism). Minimally, *a theory of mind is representational if it attempts to explain intentional concepts by appealing to the notion of mental representation.* (We will examine the notion of intentionality shortly so at this stage let us settle for 'belief', 'desire', 'intention', and 'understanding' as examples of intentional concepts.) Works of Descartes, Hume, and

Locke offered early forms of representationalism by explaining our cognitive contact with the world as being mediated by ideas or impressions. Insofar as the ideas and impressions allow the subject to represent the world to him or herself the theorist who posits their existence to explain intentional concepts is propounding a RTM (see Palmer, 1988, pp.124-125 for the connection between Descartes' 'ideas' and the notion of representation in Artificial Intelligence).[2]

Propositional attitudes, such as beliefs, desires, and the like, are *about* things, for if one has a belief then one has a belief about something. Propositional attitude Realism holds that there are such things as states of belief and that they are states of those organisms which believe. Thus, it follows that propositional attitudes states and, therefore, states of an organism, are about things; they have content. It is Fodor's opinion that this is a supposition of common-sense psychology and that, given this supposition, the best explanation of what propositional attitudes are requires the adoption of the Representational Theory of Mind (see Fodor, 1981, p.26, and 1985, p.11). One way of expressing this opinion is to say that the representational theory of mind offers a route to explaining the intentionality of propositional attitudes.

2 INTENTIONALITY

a) Intentional Objects

The epithet 'intentional' can be used to describe substantial phenomena such as states, events, or processes, or to describe linguistic phenomena such as verbs, predicates, or sentences. The Realist can adopt both uses with a clear conscience (though the first is more natural) but the Irrealist will have difficulties with the first since, with the obvious exception of

[2] There are problems with retrospectively attributing the RTM to ideational theorists such as these. For example, Locke equates ideas of objects with 'external visible resemblances' (Locke, 1690, II,.XI.17) and a resemblance relation is not the same as a representation relation. Although it follows that if A resembles B, then B resembles A such symmetry is not implied by a representation relation. There is, strictly speaking, a difference between resemblances and representations but, be that as it may, the similarities between the earlier ideational and later representational theories can be striking.

negative existential generalisations (such as 'There are no intentional states'), sentences about intentional states will usually quantify over them. That said, we need an account of the first sense of 'intentional' because this is the key to understanding why it should be thought that Realism requires the Representational Theory of Mind.

In keeping with common practice within the recent philosophy of mind I will explicate the notion of intentionality by quoting from Franz Brentano's *Psychology from an Empirical Standpoint* (1874). By way of a definition of mental phenomena Brentano offers the following:

> 'Every mental phenomenon is characterized by what the Scholastics of the Middle Ages called the intentional (or mental) inexistence of an object, and what we might call, though not wholly unambiguously, reference to a content, direction toward an object (which is not to be understood here as meaning a thing), or immanent objectivity. Every mental phenomenon includes something as object within itself, although they do not all do so in the same way...We can, therefore, define mental phenomena by saying that they are those phenomena which contain an object intentionally within themselves.' (Brentano, 1874, pp.88&89).

The suggestion that all mental phenomena are 'directed toward an object' has been questioned on the ground that, for example, sensations of pain have no such direction, but since our concern is with cognition, and in particular with propositional attitudes, such questions can be overlooked.

An alternative definition of mental phenomena offered by Brentano maintains that they characteristically present something to the subject. Indeed, he says that, 'this act of presentation forms the foundation not merely of the act of judging, but also of desiring and of every other mental act' (ibid., p.80). To elucidate his use of the noun 'presentation' Brentano points out that 'by "presentation" we do not mean that which is presented but rather the act of presenting it' (ibid.) and of the verb he says '"to present", "to be presented" means the same as "to appear"" (ibid., p.81). It would seem to follow that what is presented is what appears and its presentation is an appearance (an appearing).

So, we can characterise the desire for Cheddar cheese as an intentional state because it entails the intentional inexistence of an object. What this means is that an object (somehow related to Cheddar cheese) has an existence in intention, that is in the mind.[3] The intentional object determines what the desire is about and is its content.[4] Alternatively we can say that the desire involves the presentation of Cheddar cheese and that in being presented the cheese 'appears' before the subject as an intentional object. It is worth noting that on Brentano's definition *every* mental phenomenon is directed toward an object, so, for example, it would not be possible to have a belief which was not a belief about something.

Taking 'intentionality' as the term for the distinguishing characteristic of mental phenomena we can define it in his terms as directedness toward, reference to, or presentation of, an intentional object. It is the name of a relation, therefore, between the subject and an object of intention. There are, however, a number of problems with the definition as it stands. Firstly, the nature of the relation between the intentional object (what is presented) and the subject, to whom it is presented, is obscure. That is, it is not easy to say what the directedness, reference, or presentation amount to in this context. Secondly, some definition or characterization of the intentional object is owing, but it is not clear that such a characterization can be given without merely stipulating that it is to be an object with intentional in-existence. Such stipulation is threatened by the charge of vacuity since we can say no more of intentional in-existence than that it is possessed by an intentional object. Lastly, the individuation of the intentional object is presumably afforded by its relation to the physical object of which it is an intentional in-existence, but the nature of this relation is, so

[3] This seems the most sensible reading of the phrase 'intentional inexistence' given Brentano's footnotes on his definition (ibid., p.88).

[4] There are clearly pitfalls surrounding Brentano's definition. It would not be correct to say that the desire for Cheddar cheese is a desire for an intentional object, understood as an object which appears in consciousness, because this would imply that the

far, unclear. We might ask, for example, what is it about an intentional object involved in a desire for cheese, that relates it to cheese rather than charcoal?

It is in responding to such problems that representationalism finds an explanatory niche. By incorporating the notion of representation into Brentano's thesis we seem to acquire a suitably mundane notion for characterizing the intentional object and explaining the relation of directedness. When one desires a piece of Cheddar one is directed toward a mental representation of the thing one desires in a way analogous to that in which someone, on viewing a photograph of an E-type Jaguar, may come to have an actual car as his object of desire. Also, the relata of the representational relation are a representation and that which it represents, and the representation itself can be characterised as a picture, or sign of some sort, with qualities independent of its representational property. Thus, identification of the intentional object with a representation appears to avoid vacuity. In addition, it makes the relation between the intentional object and the physical object a familiar one, for when we think of representations as pictures, for example, the question of what is represented in the mind (cheese or charcoal) can be resolved by appeal to the notions of resemblance and pictorial context (although notions of resemblance and representation should not be equated, as we noted above).[5] The RTM, then allows us to identify, firstly, the intentional object with a representation and, secondly, the relation between an intentional object and a physical object as a representational relation.

desire for Cheddar cheese is the desire for its appearance before the mind. It would be more correct to say that the desire is for a physical object which has intentional in-existence.

[5] In cases where there would not be any existing physical object to represent, as in thoughts about unicorns, it might seem that there can be no representational relation. A standard response from those who adhere to an imagistic RTM would be to say that the mental representations are composite and the representation of a mythical creature is a composition of representations of existing parts, such as horses and horns. An alternative response, open to those for whom mental representations are sentential, is to say that the symbol 'unicorn' represents the property of being a unicorn (see Fodor, 1990, pp.100&101) or that the representation is a concept which can be broken down into subconcepts which represent existing microfeatures believed to be constituent of unicorns - a response one might receive from Smolensky.

b) Intentional Language

When propositional attitude states and processes are described as intentional, the object, or content, upon which they are supposed to be directed will be provided by the proposition which follows the word 'that' in the attitude ascription, as in 'John believes that *Cheddar is a cheese*'.[6] From the point of view of logic the important feature of propositional attitude ascriptions is that the content proposition is often brought into an *intensional* context (or a *referentially opaque* context as Quine would have it (see Quine, 1956)), that is, a context which bars quantification over the subject term of the proposition. In such a context the way the proposition is expressed is all important to the truth of the ascription of the attitude. If we explain the extension of a term as the set of all *things* of which it is true then we can see that in an extensional (referentially transparent) context the proposition 'Mohammed Ali won an Olympic gold medal' is equivalent to the proposition 'Cassius Clay won an Olympic gold medal' since the extensions of the terms 'Mohammed Ali' and 'Cassius Clay' are one and the same person. However, the intension (the sense or, loosely, the meaning) of the two terms is not the same and this is of importance in intensional contexts. Thus, although the assertion 'John believes that Cassius Clay won an Olympic gold medal' may be true, John's ignorance of Clay's subsequent name change makes the assertion 'John believes that Mohammed Ali won an Olympic gold medal' false. Introducing the propositional attitude ascription into the assertions divests the embedded propositions of equivalence.

So sentences ascribing propositional attitudes generally form intensional rather than extensional contexts. Just as important is the fact that often those sentences are not truth functional. That is, the truth of the ascription is logically independent of the truth of the embedded proposition. Thus, the truth of the sentence 'John believes that *Cassius Clay won a gold medal in the 1963 Olympics*' is independent of the truth, or (in this case) falsity, of the

[6] As far as I am aware, it was Russell who coined the phrase 'propositional attitude'. See Russell, 1940, P.18.

embedded proposition. The attitude of knowing provides an exception here since the truth of the attribution of knowledge of *p* to someone does entail that *p* is a true proposition.

It is because propositional attitude ascriptions are not truth functional that they are forbidden entry into logical relations such as contradiction and entailment. Thus, from

1) 'Socrates was a man or Socrates was a woman', and

2) 'Socrates was not a woman', we can derive

3) 'Socrates was a man',

but, from

4) 'John believes Socrates was a man or Socrates was a woman' and,

5) 'John believes Socrates was not a woman', we cannot derive

6) 'John believes Socrates was a man',

because John may not appreciate that 1) and 2) *entail* 3). That said, logical relationships do hold between the embedded propositions, that is, between 1), 2), and 3) as embedded in the ascriptions 4), 5), and 6), given above. This is why we can say that John *should* believe 3) if he believes 1) and 2). So although the relations between propositional attitudes ascriptions are not logical (because the ascriptions are not truth functional) they are *normative.* This will be important later.

It is the intensionality of propositional attitude ascriptions that allows us to characterise certain linguistic items as *intentional.* Intentional verbs, predicates and sentences are those that create or, in the case of sentences, constitute intensional contexts. However, such a characterisation is not rigorous because, for example, modal operators also create intensional contexts in which the substitution of coextensive expressions is not permissible. While it is true that eleven is necessarily greater than ten and also that I have eleven coins in my pocket, it is not true that the number of coins in my pocket is necessarily greater than ten. So some expressions which create intensional contexts are not intentional, in the sense of psychological.

Furthermore, many uses of psychological verbs (such as those relating to perception) do not create intensional contexts. For example, the sentence 'John saw that William was injured' would imply that William was indeed injured.

The intensionality of propositional attitude ascriptions provides another motivation for the Realist to adopt a Representational Theory of Mind. A standard Realism will propose that believing, for example, is an internal state relating a subject to an object of belief, and the grammatical form of a belief ascription suggests this, for '*x* believes *p*' gives 'believes' a (grammatically) transitive role which relates the subject phrase, '*x*', to an object phrase, '*p*'. In the case of an extensional context provided by a sentence like 'John held a brick' we can say that the holding was a state involving John and a brick, thereby relating a person and an object. However, when we look at the belief ascription it is difficult for us to say what the object of the belief is. If John believes that *a rat is in the shed* he is not in a relation to the spoken or written sentence 'A rat is in the shed' for it is not the sentence but what it says that he believes. However, to say that it is the proposition that John is related to will, for a Realist, entail that propositions are objects and that is not a promising premiss for a naturalistic explanation of cognition. The intensionality of the belief ascription denies the Realist the option of having the content of the proposition (the object, or state of affairs, it is about) play the role of the object of belief because the content of the proposition is not imported into the ascription, for an intensional context is one which will not permit quantification over terms in an embedded proposition.

The representationalist has a way round this problem, however, because he can claim that a representation is the object of belief. The representation will have as its content the content of the proposition it expresses and, since a representation can misrepresent, this does not require the existence of the state of affairs represented. The Realist wants belief to be an extensional relation, in this case realized within the subject, and the representational theory of mind appears to permit this by allowing quantification over representations as objects of belief.

However, we should note that the grammar of belief ascriptions does not entail this ontological commitment. In fact, the perceived requirement for an ontology of intentional objects results from a conflation of the notion of a grammatical object with that of a material object. In the context of a lesson on grammar a pupil might be given the sentence 'John believes that there is a rat in the shed' and asked 'What is the object of John's belief?'. The answer would be to recite the clause 'that there is a rat in the shed'. Clearly the question is not the same as 'What object is John related to by his belief?', where 'object' is meant materially, because a grammatical object is not a material object. If it was then John's belief about a rat would be a belief about a grammatical object. To bring out the absurdity we might consider the sentence 'John sent Mary a book' for which the answer to the question 'what is the direct object of the verb?' would be 'a book'. Now if the direct object is seen as a material object then we are led to the conclusion that what John sent Mary was a direct object and so a correct answer to the question 'What is the direct object of the verb?' would be 'a direct object'.[7]

Therefore, there is good reason to argue that it is a misunderstanding of grammar which leads some to perceive a need for an ontology of intentional objects. That need is created only by a construal of psychological verbs as referring to relational states or processes and this construal, when it is attributed to the 'folk' who use those verbs, gives rise to the assumption that the folk psychological vocabulary is fundamentally referential. This mistaken assumption is a basic tenet of both Realism and Eliminativism.

However, once the assumption has been made, the advantage of a Representational Theory of Mind is that it provides a framework within which to explain how beliefs, desires, and the

[7] This example is taken from 'The Intentionality of Sensations: A Grammatical Feature' by G.E.M. Anscombe (1965). In this paper Anscombe argues that 'intentional objects are a sub-class of direct objects' (p.6) and my argument is intended as a variation on this theme.

like, can be internal states. Given that beliefs are about something, what we required, and what Brentano tried to provide, was an account of how internal states could have this property. In effect, what we required was an account of how propositional attitude states could have semantic properties.

Semantics properties are, broadly speaking, those properties relating a sign to what it signifies. A proposition may be expressed by a sign in the form of a written sentence, a picture, or a diagram, and the semantic properties of propositional signs are those relating the sign to possible states of affairs which they represent. The properties will include reference (as the proposition will contain terms referring to objects in the state of affairs), truth (for if the state of affairs exists then the proposition will be true), and content or 'aboutness' (for the proposition's content is the state of affairs it is about and its having content gives it 'aboutness').

When the representationalist claims that a propositional attitude is a state relating a subject to a proposition, expressed by an internal representation, the internal state immediately acquires the semantic properties belonging to the representation, for representations (often) have the same properties as signs. This is why Fodor claims that 'propositional attitudes inherit their semantic properties from those of the mental representations that function as their objects' (Fodor, 1981, p.26). What is represented is the content of the proposition, that is, what it is about. What a proposition is about will be a state of affairs so we can say that the mental representation is also about a state of affairs in the sense that it refers to it or, loosely, means it. If what is represented is an actual state of affairs then one might say that it is a true representation.

So, by combining Brentano's characterisation of mental phenomena with the Representational Theory of Mind we have arrived at what I take to be the modern conception of intentionality as a property of states, processes, and events. But we should note that it is a separate conception from those involved in characterizing intentional language.[8] That said, it has allowed us to explain the Realist's view that propositional attitudes are both internal states, and states with semantic properties.

3 THE INFINITE REGRESS ARGUMENT

a) The Argument

Representationalism provides an explanation of how propositional attitudes, as internal states or processes, can be about something—it does so by making representations constituents of the attitudes—and how propositional attitudes can be states contributing to the ætiology of action. For example, we might explain why John fetched his umbrella by saying that it is because he believed that it is raining—it is what John's belief is about that accounts for his action—and the belief's 'aboutness' can be explained, via the Representational Theory of Mind, as a matter of John having a representation of the possible fact that it is raining. If we ask how this representation led John to act as he did we might expect the answer that he understood what the representation meant. However, given that representations are being employed to explain the role of propositional attitudes (in this case John's belief), and understanding is also a propositional attitude, we now need to posit a further representation to explain how John understood the representation constituent of his belief. Of course, that further representation can do no work until it too is understood, but this will require a third representation and, therefore, another act of understanding and so on ad infinitum. Thus, if, as is usually the case with representation, mental representations must be understood if they are to be acted upon *as representations*, they cannot explain the role of propositional attitudes in the causation of action.

[8] That is, one can talk of intentional verbs, predicates and sentences without commitment to either the Representational Theory of Mind, or to Brentano's characterization of intentionality.

This form of argument was used by Gilbert Ryle in 'The Concept of Mind' (1949) as a *reductio ad absurdum* of the 'intellectualist legend' according to which intelligent action involves both the performance of the action and the prior mental operation of planning the action. Since the prior mental operation can also be performed intelligently, or stupidly, it would follow that it too involves a prior operation of planning which, of course, might also be intelligent, or stupid. The intellectualist explanation of intelligent action, therefore, produces an infinite regress (ibid., pp.31&32). Ryle makes a similar objection to the 'myth of volitions' according to which voluntary action requires a prior act of will. The regress ensues because the question arises whether the act of will is, itself, voluntary. If it is not then the bodily movement it produces could not be said to be voluntary because, for instance, 'if I can not help willing to pull the trigger, it would be absurd to describe my pulling it as "voluntary"'. If the act of willing *is* voluntary then, since the 'myth' requires that voluntary action be preceded by an act of will, that act of will requires a prior one, and so too with that act, and so on ad infinitum (ibid., pp.65&66).

These objections are not aimed at representationalism but at the idea that cognitive and conative processes must precede intelligent and voluntary action. Their targets, therefore, are certain types of explanation of action and their force is acquired when the actions are divided into physical and mental components; the bodily movements and the mental acts of willing or planning. The regress argument, as I have just used it, has as its target the postulation of representations in explanations of how propositional attitudes bring about actions. In this case the force of the argument is provided by the attempt to use *representation* (which might be called a semantic concept) to account for a, purportedly, causal relationship between an intentional state and behaviour associated with it.

Both targets seem to share the requirement that there be an entity within. In the case of Ryle's target, it is an entity which *plans*, or *wills*, an action, while in our case it is an entity which brings about action because it *understands* a representation of a fact. Since these (italicized) verbs are usually applied in ascriptions made of people, the inner entity is often characterised as an inner person; a homunculus. The regress is created by the aforementioned explanations because they attempt to account for people's actions by appeal to inner mental actions which imply the existence of inner 'people' whose own mental actions will also require explanation and so on.[9] The moral, one might think, is to avoid explanations which have, or entail, recourse to inner actions or, in effect, to avoid postulating internal cognitive processes. However, cognitive scientists do not tend to accept this as a moral and, as we shall see, some believe the infinite regress can be eluded by introducing the the computer metaphor.

To the cognitive scientist the infinite regress argument may seem insidious for it lies in wait at some point on many an explanatory tramp. We noted its use by Dennett to revive the problem of allowing that we hold an infinite number of tacit beliefs when, by way of solution, it was assumed that we have an extrapolator-deducer mechanism attached to a core library of sentential representations (see chapter 2, section 2 c)). Although, in 1969, Dennett presented a version of the argument, which I will sketch shortly, his 'Artificial Intelligence as Philosophy and as Psychology' (Dennett, 1978a) offers a strategy for halting the regress. Let us turn to this.

b) Dennett's Reply

Dennett characterises Hume's problem (see note 9) as the problem of creating representations which understand themselves . This is because 'nothing is intrinsically a

[9] The idea of an inner person is associated with the idea of an inner self. The difficulties with the latter notion are familiar from the philosophical considerations surrounding the notion of personal identity a focal point of which is Hume's argument that though we have ideas and impressions of objects and events we do not have an impression of a self which has these ideas and impressions (Hume, 1740, 'Appendix to the Treatise'). The relationship between this and the regress argument is illuminated by the observation that the ideas and impressions imply, and would seem to be inert without, an agent who can associate them to form judgements, but who would also need ideas and impressions of the original ideas and impressions. Hume's argument points out that we have no acquaintance with such an agent, while the regress argument points out that an impossible infinity of agents would be needed to animate the contents of the mind.

It is because of the relationship between Hume's and the regress argument that Dennett conflates the two under the title 'Hume's Problem' (Dennett, 1978a, pp.122&123).

representation of anything; something is a representation only *for* or *to* someone' and, consequently, the existence of a representation requires the existence of a representation '*user* or *interpreter*'. In the case of mental representation the requirement is of a self or homunculus, who uses or interprets the representations as a preliminary to action, and this is the source of the infinite regress. However, according to Dennett, 'Homunculi are *bogeymen* only if they duplicate *entire* the talents they are rung in to explain' (ibid., pp.122&123).

This last observation indicates the route to the solution of Hume's problem that Dennett believes arises from the models used by researchers in Artificial Intelligence. A common method of modelling is that of constructing flow charts, and Dennett explains that,

> 'a flow chart is typically the organizational chart of a committee of homunculi (investigators, librarians, accountants, executives); each box specifies a homunculus by prescribing a function *without saying how it is to be accomplished* (one says in effect: put a little man in there to do the job). If we then look closer at the individual boxes we see that the function of each is accomplished by subdividing it via another flow chart into still smaller, more stupid homunculi. Eventually this nesting of boxes lands you with homunculi so stupid (all they have to do is remember whether to say yes or no when asked) that they can be, as one says, 'replaced by a machine'. One *discharges* fancy homunculi from one's scheme by organizing armies of idiots to do the work.' (ibid., pp.123&124).

The idea, then, is that a cognitive task, such as deducing the whereabouts of a missing object, is broken down into sub-tasks which are then broken down into further sub-tasks and so on until the bottom level sub-task is fulfilled by a binary element such as a switching mechanism. The top level specification of the task will employ intentional concepts (in this case the concept 'deduction') and, hence, imply the existence of a 'talented' homunculus, but on breaking down the task it can be seen that this homunculus can be replaced by less talented ones (who might simply have to answer questions like 'Is the object in the kitchen?'), and so on until the tasks of the bottom level homunculi can be described without use of intentional concepts.

In fact this is not quite what Dennett says, for he allows that the bottom level homunculi have to 'remember whether to say yes or no', but if he is to avoid the infinite regress he must rely on the thought that the stupid homunculi are more like switches than operators of switches.[10] An operator of a switch still has to remember to turn it on or off and *remembering*, as an intentional concept, is supposedly explicable only by postulating representations which are used by a subject (for a similar point see Palmer, 1988, p.128). In order to avoid the infinite regress argument Dennett's strategy must yield homunculi-free bottom level tasks, and this amounts to the complete removal of intentional concepts from the bottom level of explanation.

The transition from a level at which tasks are performed by homunculi to one where they are not—the transition from switch operator to switching mechanism—is the corollary of the transition from representation use to physical cause and effect. What is noteworthy about this is that if the transition can be made at the bottom level then it will reflect back up through the system. This can be appreciated if we consider that a high-level task is specified by a flow chart, and that each step of the task will correspond to a box in the flow chart which itself contains an embedded flow chart, each box of which will contain a further flow chart and so on until the last flow chart's boxes detail simple switching events or their equivalent. Thus for each high-level task there will be a nesting of tasks, corresponding to flow chart boxes, the most deeply embedded of which are simple operations which can be performed by switching mechanisms. The fact that they are embedded means that each task can be specified by the bottom-level mechanistic operations alone. It hardly needs pointing out that the switching events will be susceptible to a causal (as opposed to a regress-inviting intentional) explanation and this propensity will be bestowed upon the tasks of the more talented homunculi higher up, for we find that, when these are specified at the lower level of explanation, they are performed by arrays of switches, or their equivalent. The regress is terminated at the mechanistic bottom

[10] See Dennett, 1978b, p.102 for a suggestion that he did have such a thought in mind.

level but it could be ended sooner because the crucial step is the transition from the intentional to the causal level of explanation.[11] Talk of homunculi avoids the need to specify the activity of the system in terms of the complex mechanistic processes which actually produce the system's output, but such a specification is required if we are to explain what really occurs within the system.

At this point I wish to press the claim that Dennett's solution to 'Hume's problem' only works if the aforementioned transition from intentional to causal levels of explanation occurs and that this transition marks the elimination of the homunculi, not just at the bottom level, but throughout the system. The import of this claim is as follows: If the regress argument is avoided by insisting that homunculi are an explanatory convenience (as Dennett implies, see his 1978a, p.123),[12] and that all cognitive processing is actually performed by mechanisms which do not employ representations (on pain of regress), then the sort of cognitive systems Dennett is describing do not employ representations at all. Indeed, Dennett acknowledges this when he says, 'it is no longer obvious that psychology needs internal representations: internal pseudo representations may do just as well' (ibid., p.125). The consequence of Dennett's response to the Hume/Ryle argument is that there are no internal representations because the are no subjects to use or interpret them. That is, there is no one to whom the representations represent anything. Therefore, the response can halt the regress only by discarding representations.

So, the reason why the Representational Theory of Mind is susceptible to the infinite regress argument is the fact, noted by Dennett, that the word 'representation' ordinarily occurs

[11] The point could be put another way. If the bottom-level tasks are carried out by simple switching operations then higher-level tasks must be performed by a complex array of switching mechanisms which will have a purely causal description. Since a switch does nothing until it is operated and the introduction of a switch operator, endowed with intentional states, generates a regress, all switch operation within the system must be carried out mechanically. Thus the infinite regress is avoided by replacing intentionality with causality.

[12] For Dennett the description of sub-systems of a computer as 'homunculi' is a consequence of taking the intentional stance and it is the adoption of this stance that is convenient.

It is hard to imagine any cognitive scientist taking the notion of homunculi literally though the postulation of internal self-sufficient cognizing modules is common in the literature (see Fodor, 1983).

in contexts where it makes sense to ask for, or to, whom the representation represents. Thus we can say of the sergeant's stripes that they are intended to represent his rank *for* other soldiers and we might say that a cartographer intends to represent a topography *for* map users.[13] If the concept of representation in explanations of cognition is supposed to be the same as that just instanced, then we should be able to ask whom mental representations are for. Of course, since representations are meaningful, the question requires an answer about an entity with cognitive capacities and consequently invites the regress argument. For this reason it cannot be said that mental representations are *for* the cognitive scientists themselves. It might be tempting to say that the entities which are mental representations need not be interpreted as such by the subject in whom they occur because they are interpreted as contentful items or events by the theorist studying the cognitive system. However, quite apart from the fact that it would now be unclear how they play a role, *qua* contentful items, in producing behaviour, the infinite regress will still ensue. If interpretation, a cognitive activity, is to be explained using the notion of representation then the cognitive scientist's interpretation of a subject's inner state as representational will entail that *she* has an internal item which can be called a representation only if it is interpreted by (that is, only if it is *for*) another cognitive scientist who now has an internal item.....

Note that it would not be permissible to say that all that is required for an internal state to be representational is that it *could* be interpreted as representing, for then, given appropriate stipulations, any state would be representational and would represent just about anything. What makes something a representation is that it is being, or has been, used as a representation, not that it could be used as one, and what determines what it is a

[13] There may appear to be cases of unintended representation. For example, a bricklayer may spend a day constructing a garden wall, and we might say that the wall represents a day of his labour although it would not be the case that the bricklayer intended the wall to represent anything. However, though this is a legitimate use of the word 'representation' the fact that there is no possibility of the wall, or the bricklayer, misrepresenting shows it is not being used in the sense required by the Realist cognitive scientist; that is, the normative sense. We will consider the Normativity Requirement in the next chapter.

representation of is how it is being, or has been, used. Furthermore, interpreting a representation is often a matter of deciding on how it is, or was, intended to be used, which sits uncomfortably with the assumed inaccessibility of internal representations to the subject's consciousness (for they cannot intend anything of that of which they know nothing).

The requirement of an infinite number of homunculi has been avoided at the cost of an infinite number of cognitive scientists (see Sayre, 1987, pp.238&239 for a similar point). Thus if the representationalist is to defuse the argument her usage of 'representation' must be divorced from what I have suggested is the ordinary usage. Specifically, those who propound the theory must insist that some things are intrinsically representational, that is, they do not represent for, or to, anything.

c) Searle's Reply

Dennett seems to recognise that the use of 'representation' is problematical but suggests that if one rescinds the concept from cognitive science 'one is in danger of discarding one of the most promising conceptual advances ever to fall into philosophers' hands' (Dennett, 1978b, p.102). My view is that the use of the word, by cognitivists, is confused and ultimately incoherent, but the infinite regress argument alone will not show this to be so. The cognitivist can argue that not all representations are understood, and that mental representations, in particular, are not.

This is the tactic employed by John Searle in his book *Intentionality* where, as a response to the infinite regress argument he says,

> 'the false premise in the argument in short is the one that says that in order for there to be a representation there must be some agent who *uses* some entity as a representation. This is true of pictures and sentences, i.e., of derived Intentionality, but not of Intentional states.'(Searle, 1983, p.22)

Searle explains that though, in the case of pictures, we can distinguish between the entity and its representational content we cannot do so in the case of intentional states and their content. We are told that 'the content determines its conditions of satisfaction' such that, for example, if the state were a belief that it is raining then its representational content will determine that the belief is satisfied if and only if it is raining. However, to be conscious of the conditions of satisfaction of ones conscious beliefs is not to have second order intentional states about ones first order beliefs (which, as Searle notes, would lead to an infinite regress), 'rather, the consciousness of the conditions of satisfaction is part of the conscious belief or the desire, since the Intentional content is internal to the states in question' (ibid.). Searle also tells us that the intentionality of utterances is *derived* from the intentionality of intentional states (ibid., pp.27&28) but also that an intentional state, such as a belief, 'does not require some outside Intentionality in order to become a representation, because if it is a belief it already *intrinsically* is a representation' (ibid., p.22, my italics).

I will not attempt a critique of Searle's theory of intentionality but will remark on his use of the word 'representation'. He maintains that 'Intentional states represent objects and states of affairs in the same sense of 'represent' that speech acts represent objects and states of affairs' though he notes that the former have intrinsic, and the latter derived, intentionality (ibid., pp.6&12). In this way he employs the notion of representation to endow intentional states with content—the intentional state inherits its 'aboutness' from the representation—while denying that this content is derived in the way that the content of a picture might be (see ibid., pp.11&12). As a representation, a speech act gets its meaning (it determines its conditions of satisfaction, to allow Searle his analysis) because it is understood in a certain way, for, as we have been told, its intentionality is derived. The question one wants to ask is what is it for an intentional state, a belief for example, to represent intrinsically?

It appears that Searle's answer is to say of the belief, 'it simply consists in an Intentional content and a psychological mode' (ibid., p.22) (a psychological mode is the attitude to the content such as the *belief* or *wish* that *p*). However, what we want to know is how a *structure* represents intrinsically, that is, how it acquires meaning or content.[14] To be told that it does so by determining its conditions of satisfaction just postpones the agony. Now we want to know how the structure determines its conditions of satisfaction (for surely having such conditions depends on the structure being about something) and it is far from clear how Searle can answer without use of concepts like interpretation, or understanding, that is, without appeal to what he calls 'derived intentionality'.

It is difficult to make sense of the idea of a structure determining its own meaning. It has certain ramifications which are deeply puzzling. For example, it entails that, at various locations in the universe, there may be structures which are representational despite the fact that no one ever has been, or will be, aware of them. Searle, no doubt, would preclude such a possibility, on the grounds that an intrinsically representational structure must have a neurobiological nature, but only by begging questions, not just about such an apparently arbitrary form of physicalism, but also as to what neurobiology has to do with meaning. The idea that intentional content is determined by the structure which has it also entails that meanings are in the head which Searle acknowledges (ibid., p.200). I shall not rehearse any standard Externalist arguments against this consequence (see, for example, Putnam, 1975) but will have something to say about the related notion of context-independence (which is required by the Internalist, but also by some Externalists) later. For now, however, we need to look at Fodor's response to the infinite regress argument.

[14] Searle suggests that 'Intentional states are both caused by and realized in the *structure* of the brain' (ibid., P.15, my italics).

d) Fodor's Reply

Fodor insists that propositional attitude states inherit their semantic properties from the internal representations partly constituting them. His claim is 'precisely that the objects of propositional attitudes are symbols (specifically mental representations) and that this fact accounts for their intensionality and semanticity' (Fodor, 1981, p.24). The infinite regress argument is, in general, directed upon the rationalization of semantic properties by appeal to internal representations, and gains its force from the suggestion that a representation's possession of semantic properties is dependant upon its being interpreted, or understood, as possessing them (or upon its having derived intentionality, to use Searle's term). Here, then, are two forms it may take when specifically aimed at Fodor's thesis:

1) As Dennett argued, the postulation of a language of syntactically analysable events or structures, or 'brain-word tokens', as a means of explaining storage in memory, would seem to require that there be mechanisms for 'reading' and 'understanding' this language which must also have syntactically analysable parts (Dennett, 1969, p.87).

2) Fodor's explanation of understanding a sentence involves its translation into a formula of the LOT but this formula will in turn need to be understood through a further process of translation and so on.

Here, in part, is Fodor's reply to 1);

> 'The argument is fundamentally wrong-headed since it assumes a picture of the nervous system as issuing commands which must be 'read' and translated into actions...by some *further* system that intervenes between the efferent nerves and the effectors. But this picture is no part of the theory. On the contrary, what is required is just that the *causal* properties of such physical events as are interpreted as messages in the internal code must be compatible with the *linguistic* properties that the interpretation assigns to those events' (Fodor, 1975, p.74 note 14).

In reply to argument 2) Fodor makes full use of the computer metaphor in suggesting that the LOT is the human equivalent of the machine language of a computer while the brain's translation mechanism is comparable to the computer's compiler (ibid., pp.65&66, and see above chapter 1 section 3 b)). A computer does not need to translate its machine language because it is built to use it and 'the machine language differs from the input/output language in that its formulae correspond directly to computationally relevant physical states and operations of the machine' (ibid., p.66). Thus the translation of a natural language sentence into a formula of the LOT is a reproduction of that sentence as a physical state or operation of the brain. The objection that this does not explain what understanding is because the formula of the LOT now has to be understood is countered by the assertion that the LOT formula, like the machine language formula, is not in need of translation because it effects an output by virtue of its causal, rather than semantic, properties. Understanding a natural language sentence requires that it be translated into an LOT formula but the regress stops here. The causal properties of the LOT formula now become explanatorily relevant to any story of how the act of understanding may modify behaviour.

Put in a way clearly relevant to the regress argument as directed at the Representational Theory of Mind, what Fodor's replies amount to is this: Mental representations, understood as physical items which are interpreted as messages in the LOT, have both causal and semantic properties and it is by virtue of the former that they are translated into actions 'or, anyhow, into muscle contractions' (ibid., p.74, note 14).[15] So, in answering the question why John's representation of the possible fact that it is raining led him to fetch an umbrella, there is no need to appeal to his understanding of the meaning of the representation (that is, there is no appeal to its semantic properties) because the action is brought about by the representation's causal properties (which it has by virtue of being a physical event). The propositional attitude, in which the representation figures, is now efficacious with respect to bodily movement not

[15] The distinction between 'muscle contractions' and actions is important. Some, who object to cognitivism, argue that while causal properties and relations within an organism might explain the ætiology of bodily movement, it will not explain action, for explanation of action presupposes a normative setting which cannot be captured by causal explanation (see, for example, Ryle, 1949 and Winch, 1960). While I think this objection is apposite it is not without respondents, most notably Donald Davidson

because of the semantic properties of the constituent representation (which would seem to require an inner, regress inviting, act of understanding) but because of the causal properties of its physical instantiation. Indeed, since Fodor has specifically denied that the semantic properties of mental representations play a part in cognitive processing he does not need a homunculi to interpret them in order to initiate action (see Fodor, 1980, p.231).

Although Fodor thinks that by using the computer metaphor he has halted the regress at the point at which content is repesented in the LOT, he appears to have missed the fact that in order for us to compare the LOT to a machine language, or code, we need to allow that there are physical (presumably neural) events that can be interpreted as messages in the internal code (see his reply to Dennett quoted above). Two observations should be made regarding this.

Firstly, the attribution of a machine *language* to a computer is metaphorical. Insofar as language is a normative activity, a machine cannot literally be said to have a language because, although it may act in accordance with a rule, a machine could not be said to follow, or be guided by, a rule (I will provide support for this contention in chapter 6). The computer metaphor relies on the idea that if a computer is a language using mechanism then there is no *a priori* impediment to assuming that cognitive activity in humans is nothing more, or less, than the mechanistic manipulation of linguistic items (*viz.*, representations). However, if what I have just said is correct then the computer metaphor itself relies on a prior metaphor derived from the very domain which it (the computer metaphor) is intended to illuminate. Using the computer to explain cognition is like pointing to the sullen sky to explain what it is for someone to be in a sullen mood.

Secondly, the interpretation of the brain's states as LOT formulae is integral to Fodor's Realistic explanation of the propositional attitudes. Unless they are interpreted as such there is no clear sense to the supposition that brain states can be attributed content or semantic

(see Davidson, 1963). The arguments I proffer in chapter 6, though not explicitly directed towards this debate, indirectly support the objection.

properties. Without these attributions there is little sense in maintaining that brain states can instantiate propositional attitudes (which is what Fodor's Realism requires), or that there is an LOT. As we have noted, the need for an interpreter of brain states *as* LOT formulae, though it need not generate an infinite regress of interpreters *within* the brain, *will* necessitate an interpreting agent, or a subject, and it is the explanation of what it is to be such an agent, or subject, for which the LOT hypothesis was introduced. The very supposition that there is a LOT will, it seems, be open to the charge of vicious circularity and the version of the infinite regress argument given above. A satisfactory account of cognition cannot attribute representational states to a system on the grounds that there is a further system interpreting the states by virtue of its own representational states.

So, what reason might Fodor have for deeming states of brains to be representational given that the answer 'because they are *interpreted* as such' will generate an infinite regress? The answer to this question is not to be found in *The Language of Thought* where he writes,

> 'It remains an open question whether internal representation, so construed, is sufficiently like natural language representation so that both can be called representation "in the same sense". But I find it hard to care much how this question should be answered' (ibid., p.78).

My point is that since it is far from clear that there is any coherent notion of internal representation, Fodor can not afford this indifference. In fact, there is reason to believe that natural language representation has features incompatible with the notion of internal representation.

In indicating the incompatibility we might begin by noting that the sense in which a sentence of a natural language is a representation is one in which it is taken to impart information about something.[16] If, for example, I return home to find a note from my wife reading 'The car will not start', my wife will have informed me of an irksome state of affairs. I

[16] Of course, not all sentence types are representational in this sense. Questions, commands, rebukes, and requests, to name but a few, are rarely intended to impart information.

can use the information as a basis for various deliberations and decisions about my plans and my use of the information shows that I have understood it. It is important to note that I would interpret the sentence only insofar as I might decide she means that the battery is flat or some such thing: I would not need to *translate* the sentence.

Representational status can be conferred upon the sentence only insofar as it is used as a representation or is taken to impart information: One might say that for something to be called a representation it must *function* as a representation. The function of my wife's message is characterized in terms of her intention to inform and of my use of the information in decision making and inferences as to the best plan of action in the light of the information. Thus the function of the sentence, that is, its role as a representation, is explained from within the intentional idiom, for this characterization employs no less than four nouns (intention, decision, inference, and plan) which, in their verb forms will render a sentential context intensional. Consequently it is not open to a theorist who posits internal representations to assert that because internal states have functional roles corresponding to the functional roles of natural language representations, they too are representational. The assertion would invite the regress argument. Regarding the open question Fodor refers to, therefore, it would seem that internal representations had better not be representations in the sense in which natural language representations are if they are to have any utility in explanations of cognition. Of course, now we might wonder in what sense LOT formulæ are representations.

The problems we have been considering arise for Fodor and his use of the Representational Theory of Mind because he relies on the notion of interpretation in justifying the claim that some events and states of the brain are representations (or messages in the LOT). If a rationale could be given for assuming such events and states have semantic content which does not employ intentional concepts then the problem will have been evaded. Fodor seems to have appreciated something like this point when he wrote;

> ‘The worry about representation is above all that the semantic (and/or the intentional) will prove permanently recalcitrant to integration in the natural order; for example, that the semantic/intentional properties of things will fail to supervene upon their physical properties. What is required to relieve the worry is therefore, at a minimum, the framing of *naturalistic* conditions for representation’ (Fodor, 1984, p.32).

In later work Fodor set about meeting such a requirement in the form of a naturalized theory of mental content. Indeed much of the recent work undertaken by philosopher cognitive scientists has been concerned with the provision and discussion of such theories. It is to this rationale that we will now turn.

CHAPTER FIVE

REPRESENTATIONAL CONTENT AND CAUSATION

The programme for naturalizing content is of utmost importance to the reduction of psychological to causal explanation, and that reduction could, justifiably, be identified as the general project of cognitive science. By 'causal explanation' I mean explanation in terms of efficient causation amongst events, states and processes. Those particulars entering into a causal relationship can be individuated independently of that causal relationship but, following Donald Davidson, we need not insist that they are.[1] For example, if one explained that the throwing of a brick caused the smashing of the window then it would follow, given this sense of 'causation', that both the antecedent and consequent events can be individuated by a description other than that used in the explanation. Since the smashing of the window might be redescribed as (amongst other things) a molecular level event, and the same goes for the throwing of the brick, this would count as a genuine causal explanation. In addition, descriptions of causal relations will be extensional because different, but co-extensive expressions, can be substituted into the descriptions without any change in their truth value.

On the version of the Representational Theory of Mind under discussion mental representations are assumed to have both semantic and causal properties. Fodor tried to exploit this assumption to evade the infinite regress but may be taken to have failed largely because the assumption that inner states have semantic properties could not be sustained without recourse to non-naturalized intentional concepts like 'interpretation' and 'understanding'. Dennett, as I presented him, attempted to avoid the infinite regress by showing that the supposed semantic properties of computational states were, in fact, causal all the time, but in

[1] In 'Actions, Reasons, and Causes', for example, Davidson argues that cause and effect can be described in such a way that they are not individuated independently of each other (see 1963, p.14). This is possible because of his ontology of events.

doing so cast doubt on the very use of the notion of inner representation in cognitive science. The aim of a naturalized theory of content, then, is to show that brain states *can* be representational because they can have semantic properties, and therefore meaning, whilst also showing that the acquisition of these properties can be explained without straying into the intentional idiom or presupposing the existence of states with content.

One way of initiating this enterprise is to identify an information bearing property of natural language representation and argue both that physical and, therefore, extensional states and events can also bear information, and that they do so by virtue of their causal relations. This is the rationale behind information based semantics.

1 INFORMATION AND NATURAL MEANING

There are cases in which physical phenomena might be said to carry information about the world. For example, since water does not flow uphill, a northerly flowing stream *means* that there is a downward gradient in that direction; shadows to the east *mean* that the sun is in the west; and, a sudden force on the passengers in one direction *means* that the train has accelerated in the opposite direction. These are examples of natural indicators, or signs, and in each case it is a causal connection between the indicator and what it indicates which allows the former to carry information about, or mean, what it indicates. Thus it is a downward gradient which *causes* the stream to flow in a particular direction, the relative position of the sun which *causes* shadows to fall toward a particular direction, and the direction of acceleration of a train which *causes* the force on its contents in the opposite direction.[2,3]

[2] Of course, it need not always be that the effect indicates its cause. For example, we might say that clouds mean rain.

[3] It is not necessary to go into the question of whether these causal relations are nomic (that is, whether they can be subsumed under a causal law) or whether the relations are simply instances of causal regularities. All that matters in the following discussion is that they are causal in the sense I have outlined above. That said, it does seem that the causal accounts we shall deal with below view causation as nomological. This implies that for S to be a natural sign of P it must be the case that 'S causes P in conditions C', or 'P causes S in C', will describe a lawlike connection so that when S does not cause P or vice versa, we may assume that C is not fulfilled.

The connection between the indicator and what it indicates, since it is a causal one, can be described without recourse to intentional concepts for, as I have suggested, causal connections are susceptible to explanation in the extensional idiom. It is, perhaps, for this reason that Dretske (from whom the examples above are taken) says that, 'the power of these events or conditions to mean what they do is independent of the way we interpret them—or, indeed, of whether we interpret or recognize them at all' (Dretske, 1986, p.158). If this is right then it follows that a physical event can mean something just because it enters into causal relations and not because it is interpreted as meaning what it does.

Once this account is adapted so that the causal connections are between neural phenomena and distal objects and events then it becomes permissible to say that neural phenomena can carry information about, or mean, those objects and events. If, for example, a brain state can be identified, via some causal pathway passing through sensory apparatus, as an effect of a distal cause such as a cow, then it seems permissible to say that the brain state has the content *cow*; that is, it means *cow*. This is the rationale behind causal theories of mental content and since it implies that, at least some, inner phenomena mean or signify things without the need of interpretation, it is a rationale which evades the infinite regress argument.

However, as it stands the rationale is insufficient as an account of how neural states can acquire representational content. It is an essential requirement for the ascription of representational content to something that whatever has it can misrepresent. We may call this the *Normativity Requirement* because normativity is a property attributable only to those representations which can represent truly, falsely, correctly, or incorrectly, and a representation misrepresents when it is false or incorrect. The notion of content required by the Realist cognitive scientist must fulfil the requirement. This is because he or she is aiming at an account of propositional attitudes which have satisfaction conditions determined by their propositional content and the content will remain the same whether or not these conditions are satisfied. Thus, when I believe that it is raining the belief is satisfied just in case it is raining. If it is not raining then the conditions are not satisfied, but it is still true that I believe it is raining. When the Representational Theory of Mind is employed to explain how my belief state has content the implication is that when the belief is unsatisfied the representation, which bears the content of the belief, misrepresents how things are in the world. The problem for information based semantics is that natural meaning is not normative.

Dretske points out that though natural signs can represent they cannot misrepresent and that is why natural meaning alone cannot be the same as representational meaning (ibid., p.159). If, for example, the force on the train's passengers is caused by the braking of the train, rather than its acceleration, then we do not say the force has misrepresented the train's acceleration. Nor do we say, when they are caused by the reflection of the sun's light off a window, that the shadows to the east misrepresent the position of the sun. In other words, a natural sign means that *P* only if *P*, and if the sign occurs while *P* does not then we say that it is not a sign of *P* rather than that it misrepresents. Thus natural meaning is not normative because a natural sign is never wrong or incorrect. If the usual causal relationship, upon which the signalling relation is founded, is not instantiated then nor is the representational relationship.

Before we examine the attempts of the causal theorists at injecting normativity into natural meaning we should note one of the problems arising from conflating natural meaning with representational meaning. As I have suggested information based semantics and, more generally, causal theories of content are grounded in the assumption that a causal connection between events is sufficient for one of the events to mean, indicate, or carry information about, the other. This assumption is, supposedly, justified by the precedent of natural meaning. Now, since it is not necessary that natural signs be recognized or interpreted but only that they occur in regular causal connections with what they signify, it follows not only that, for example, the sun in the west means the shadows are in the east, the downward gradient towards the north means that the stream will flow in that direction and so on, but that any relation instantiating a causal regularity or law will yield meaningful relata. The consequence is what Fodor calls 'Pansemanticism', that is, 'the idea that meaning is just *everywhere*', and he suggests this 'is a natural conclusion to draw from informational analyses of content'. He objects to the idea because it rests on the intuition that 'means' is univocal and means 'carries information about'.[4] That this is wrong is clear, for if we accept that 'means' means the same in 'smoke means fire' and '"smoke" means *smoke*' for since 'carries information about' is transitive it would follow that 'smoke' means *fire* which is false (see Fodor, 1990a, pp.92&93).

2 CAUSAL THEORIES AND THE NORMATIVITY REQUIREMENT

Although Fodor rejects the conflation of 'meaning' with 'carrying information about' he does adopt the thesis at the core of information semantics, *viz.* that internal states represent their causes. However since, as it stands, this thesis does not satisfy the normativity requirement he has offered ingenious suggestions as to how it can be modified to do so. His most recent attempt, and one of the most successful causal theories to date, is the Asymmetric Dependence theory of content.

Prior to modification a causal theory of content might explain how a brain state can have content as follows: The mental symbol 'cow' means *cow* if cows are nomically (lawfully) related to 'cow' tokens (as a consequence of the fact that cows are invariably *causes* of 'cow' symbol tokens). That is, we have the brain state symbolizing as a natural sign for the presence of cows. Of course, natural signs do not misrepresent, so even if the symbol token 'cow' is occurrent when there are no cows present (just a horse on a dark night) we should not say that 'cow' misrepresents.

[4] One would like to think that the consequence that just about any event is meaningful whether or not anybody knows of the event, let alone what it means, is sufficient reason to reject Pansemanticism.

It might be thought that the reason natural signs do not misrepresent is that they represent whatever they are caused by. Thus, although cows are sufficient causes of 'cow' symbols, when a horse on a dark night is the cause of a 'cow' token that symbol token is not an indicator of a cow but of a horse. Thus we would have to say that 'cow' symbols represent *cow or horse* . In this case the meaning of 'cow' is disjunctive. The result of this reasoning is still that natural meaning does not allow for misrepresentation since whatever is nomically related to the sign is part of its meaning. The Normativity Requirement is met, therefore, when the, so called, Disjunction Problem is solved.

Fodor's Asymmetric Dependence theory is intended to rectify the lack of normativity by restricting the type of causal relation sufficient for content as follows: 'Cow' means *cow* if, a) there is a nomic relation between the property of being a cow and the property of being a cause of 'cow' tokens and, b) when there are nomic relations between other properties (for example, being a horse on a dark night) and the property of being a cause of 'cow' tokens then these nomic relations are asymmetrically dependent upon the original one. Thus the thesis can be explained by the counterfactual; if there were no cow caused 'cow' tokens there would not be any non-cow caused 'cow' tokens, but not the converse. A possible world analysis of the counterfactual would yield: In nearby possible worlds in which there are no non-cow caused 'cow' tokens there are cow caused 'cow' tokens, but in nearby possible worlds in which there are no cow caused 'cow' tokens there cannot be non-cow caused 'cow' tokens (see Fodor, 1990a, pp.90ff).

Misrepresentation thus occurs when the nomic relation instantiated by a 'cow' token and its cause is asymmetrically dependent on the nomic relation between cows and 'cow's (I have dropped the 'property' phrasing for the sake of simplicity—at this point it need not concern us here whether it is cows that cause 'cow' tokens or the property of being a cow). Thus, if a horse on a dark night causes a 'cow' token then the token misrepresents.

The mental symbol ‘cow’ does not mean *cow or horse* because mental symbols do not represent the asymmetrically dependent causes of their tokenings, and the relation of horses to 'cow', though nomic (because they invariably cause ‘cow’ tokens in suitable circumstances), is dependent upon the relation of cows to ‘cow’ tokens but not vice versa. Fodor's theory, therefore, seems to have solved the disjunction problem. It is highly relevant to the current examination of the Representational Theory of Mind because it seems to provide an adequate account of internal representation (since it meets the Normativity Requirement of explaining misrepresentation) which is wholly naturalistic, for it yields the conclusion that ‘representation is just a certain kind of causal relation—it's just information plus asymmetric dependence...’ (ibid., p.129). The fact that it is a naturalistic account of representation is what gives it immunity from the infinite regress argument.

It is important to note that although, when caused by a non-cow, the 'cow' token misrepresents, it can do so only because it still means *cow*. What it would misrepresent is, presumably, the fact that a cow is present, but in order to do so it must retain its representational content, namely *cow*, which it acquires by entering into nomic relations with cows which are not dependent on any other nomic relations. There are parallels between this strategy for fixing the representational content of a mental symbol and the strategies employed by other theorists attempting to naturalize content. The latter tend to do this by stipulating the type of situation in which the cause of a symbol defines, or fixes, its content. Thus, for example, one strategy (see Dretske, 1981, pp.193-195) is to stipulate that the context fixing situation is the one in which a symbol is learnt, as this situation is biased towards the causes of the symbol being those about which the symbol is to carry information. When tokens of this symbol type are triggered by causes of a type other than the one that caused the symbol tokens in the learning situation, then the subsequent tokens misrepresent. A second strategy, proposed by Ruth Millikan, is to stipulate that the content-fixing situations are the ones in which the cognitive mechanisms which produce, or use, the symbol are in compliance with their normal, or preferably, *proper* function. The proper function of a cognitive mechanism is supposed to be, more or less, that function which has been selected for its contribution to the survival of the organism possessing that mechanism.

The second strategy is the hallmark of Teleological theories of content which specify the requirements for content-fixation by appeal to the function of content-fixing mechanisms. There would appear to be two ways in which teleological appeals to function can determine representational content. Either content-fixation can depend on the functioning of the mechanism which mediates between a distal stimuli and the representation it causes[5], or it can depend on the function of the mechanism which uses the brain state as a representation of the distal stimuli. The distinction is noteworthy for two reasons. Firstly, the latter account is not a causal/informational account of content since the content of the inner representation is determined by the proper functioning of the mechanism which uses it and not by the cause of the inner state (see Millikan, 1989, pp.247-251).[6] Secondly, the latter account, in its postulation of representation users, is vulnerable to the infinite regress argument, and the charge of circularity, when it is used to explain cognitive concepts. When Millikan tells us that ‘the part of the system which consumes representations must *understand* the representations proffered to it’ (ibid., p.246, my italics) she cannot account for the concept of *understanding*, or any similar concept, by appeal to these representations.

The story told by Millikan is supposedly preferable to the former, causal, account which has been criticised on the grounds that it allows there to be indeterminate contents (and, hence, it fails to solve the disjunction problem by showing how misrepresentation is possible) by both

[5] Fodor made such an appeal to teleology at one point (see Fodor, 1990a, particularly pp.324-326) but soon dismissed it.

[6] Millikan explains the relation between representation producing and using mechanisms as follows:

> 'Although a representation always is something that is produced by a system whose proper function is to make that representation correspond by rule to the world, what the rule of correspondence is, what gives definition to this function, is determined entirely by the representation's consumers'. (Millikan, 1989, p.247).

Thus the causal relation mediated by the representation producer is insufficient for content-fixation.

Dretske (1986) and Fodor (1990, Chapter 3) (see also Millikan, 1991). Although Millikan's story may remain outside the scope of objections to the causal account it does so only because it is not wholly naturalistic and my insistence is that only a naturalistic account of what it is for an inner state to have representational content can evade the infinite regress argument and the charge of circularity.

We must now attend to the question of whether causal theories of content can, in principle, meet the Normativity Requirement, that is, whether these accounts can offer a naturalistic justification for the claim that certain brain states have representational content. Such a justification is recognized by the causal theorists as a necessary condition for assigning of meaning to inner states. Without it there can be no identification of inner states as representations. In what follows I shall treat Fodor's asymmetric dependence account as the flagship of the causal theorists' fleet.

3 REPRESENTATIONAL CONTENT

The Normativity Requirement arises, as we have seen, because the Representational Theory of Mind must allow for the fact that there can be false beliefs. In other words, if it is assumed that having a belief that *p* is being in a relation to a representation of *p* then having a false belief that *p* must be a matter of being related to a false representation of *p*, that is, a misrepresentation. Beliefs have normativity by virtue of their being about something—their intentionality, as it is called in the literature—for it is this that allows them to be judged to be true or false. However, according to the Representational Theory of Mind, this normativity is inherited from that of the inner representations which carry the beliefs' content. What is at issue, then, is whether causal theories give a coherent account of normativity at the level of inner representations.

Now Fodor tells us that a 'cow' symbol token has representational, and thus normative, content if it has a cause with an asymmetric dependence base. However, this makes little sense because a 'cow' symbol is neither true nor false, correct nor incorrect when it occurs in isolation. In Fregean terms we might say that it lacks a judgeable content or, in other words, the content of the symbol is not a content of a possible judgement. That is, we could not judge the symbol true or false any more than we might judge the written word 'cow', isolated from any context of communication, to be true or false. The word 'cow', isolated from any context, means (normatively) nothing for it expresses nothing that can be judged true or false (see Frege, 1879, p.2, for a related remark about the word 'house').

Fodor's idea is that 'symbol types express the property whose instantiations reliably cause their tokenings' (Fodor, 1987, pp.99&101 and see pp.107&108) but this will allay our objection only if expressions of properties are true or false, which they are not. If we allow that 'cow' expresses the property of being a cow, then 'cow' means *being a cow*, and it is clear that the expression alone is not true or false. Fodor also explains that according to the causal theory of content 'symbol tokenings denote their causes' (ibid.). If the sense in which the symbols denote is the same as that in which names do then, again, the symbols are not true or false. A word ordinarily used as a name, like 'Vienna', can have a truth value only insofar as it constitutes an elliptical sentence, as when it is offered in reply to a question like 'What is the capital of Austria?'. In such a context we can say that the name 'Vienna' abbreviates the proposition 'Vienna is the capital of Austria', but such contextual provisions are not available in the case of internal symbols; at least, not without the risk of introducing regress inviting homunculi to pose and answer questions. Besides, Fodor's contention is that information plus asymmetric dependence is sufficient for content and not information plus asymmetric dependence plus contextual provisions.

Of course, normativity does not apply only in areas where there can be truth and falsity. There are rules in sport and social etiquette, for example, which codify correctness, but not truth, in proceedings. In polite society, for instance, it is sometimes more correct to tell a lie than an offensive truth. More relevantly, we will often say of a word used to denote an object that it is incorrect rather than false, as when a trainee electrician identifies a resistor as a 'capacitor'. However, in such a case what is incorrect is the trainee's *application* of the denoting expression and this is important. For us to agree that a 'cow' token misrepresents we will need to attribute the normative notions of either falsity or incorrectness to the symbol token. Since the symbol is denotational it is not true or false in isolation but nor is it correct or incorrect since it has not been *applied* to anything. Both aspects of normativity can be introduced only when the symbol is put into a context. In the case of truth this context will be, at a minimum, such that the symbol occurs within, or is understood as expressing, a proposition. In the case of corrigibility, the context will involve the application of the symbol although, again, this will often be such that the symbol occurs within, or is understood as expressing, a proposition. When the property of being a cow causes a certain brain state token that token is not being applied to the cow. Referring to that brain state as a 'cow' symbol token does not change this fact. Nor does the claim that the cow - to - 'cow' relation will support certain counterfactual generalisations.

The point is that it is the propositions embedded in propositional attitude ascriptions that are true or false and, although the 'cow' symbol might be supposed to be part of a Mentalese sentence expressing a proposition, it does not, by itself, express one. The disputes among causal theorists over one another's claims to have solved the Disjunction problem (to have provided an account of why 'cow' means *cow* and not *cow or horse*) are misconceived because it is a mistake to suppose that a denoting expression is true or false, or that it is correct or incorrect, when it is not applied. This conclusion may be disconcerting to those causal theorists who claim to have found a way of demonstrating that some mental symbols can misrepresent but it is not clear that it is an obstacle to the naturalization programme in which they are engaged. In fact, if the Normativity Requirement no longer constrains the project of fixing the content of, what Fodor calls, 'items in the primitive nonlogical vocabulary of the language to which the [mental] symbols belong' (ibid., p.98) then so much the better for naturalism.[7]

Of course, reference to normativity *is* made when explicating the role of denoting expressions because, as we noted, their application is corrigible. That is, a denotational expression can be applied correctly or incorrectly. The question which must be addressed, therefore, is 'What is it to apply an internal denoting expression?'

It may be that a cogent answer to this question will emerge from a marrying of the Representational Theory of Mind to the Language of Thought hypothesis. This will allow the proposal that internal denoting expressions are applied when they combine with other items of the inner lexicon to form representations with propositional content. This answer becomes apposite when we remember that a motivation for adopting a Representational Theory of Mind is to justify the Realist claim that there are inner states encoding the content of propositional attitudes. Since the Realist thesis requires that there be inner representations with propositional content, and the formulation of a proposition might be said to require the *application* of a subject and predicate term, it would be plausible to claim that representations which encode the propositional contents of the attitudes are created by combining elements of the inner lexicon.

Clearly, the Language of Thought hypothesis provides an explanatory schema with which to account for the creation of propositional, and judgeable, contents from items of an inner

[7] Regarding items of the logical vocabulary, Fodor is happy to accept an account of their meaning in terms of functional role. Thus, 'a speaker means and by "and" iff, ceteris paribus, he has "P and Q" in his belief box iff he has "P" in his belief box and he has "Q" in his belief box' (Fodor, 1990, pp. 110&111). This implies that, in fact, the logical vocabulary does not consist of items but of relations between items of the non-logical vocabulary. Since the relations in question will have to be causal, the

lexicon. The schema may, thus, be taken to solve two problems. Firstly, it makes more palatable the claim that brain states have the property of being denotational. After all, denoting expressions are applied, and their application can be evaluated as correct or incorrect. Secondly, it appears to offer a route to satisfying the Normativity Requirement because it provides an explanation of how inner representational structures can misrepresent, via an account of how symbol elements combine to express false propositions. In chapter 6 I will give grounds for believing that 'correct application' is not an evaluation applicable to brain activity. In the rest of this chapter I will attempt to show why internal propositions cannot be constructed.

Before moving on we should note that not all propositional attitude Realists accept either the Representational Theory of Mind or the Language of Thought hypothesis. For example, some see the propositional attitudes as monadic states which can be individuated because of an isomorphism between the causal roles of the states and the inferential roles which exist among the attitudes, by virtue of their content, and between the attitudes, behaviour, and stimuli.[8] I shall not engage with this view here, however, save to say that those who take it will depend, for the correct individuation of the monadic states, upon a syncretism of causal and inferential relations and that syncretism will be the target of the arguments of chapter 6. Fodor, of course, insists that propositional attitudes are polyadic because they consist of a relation between representations and an organism. As he sees it, it is only when one conjoins the Representational Theory of Mind and the Language of Thought hypothesis that one can offer an account of the systematicity, productivity, and inferential coherence of thought and language (see Chapter 3, section 2 a)).

objections I offer in the next chapter, that there is incoherence in the attempt to explain the normativity of language use by appeal to causation, will encompass Fodor's conscription of functional role semantics for the logical vocabulary.

[8] See Fodor (1985, pp.13-15). Although Fodor does not give any examples of adherents to Monadic Functionalism, I take him to be alluding to the Turing-Machine Functionalism of Putnam and to others, such as David Lewis and Robert Stalnaker, who abjure Representationalism because they see no need for internal content bearers when functional description will afford the individuation of intentional states. Of course, as it stands the monadic view leaves unanswered the question of how functional states acquire content. Dennett, a descendant of Turing-Machine Functionalism, takes the easy way out by settling for derived

4 CONSTRUCTING JUDGEABLE CONTENTS

a) Compositional Semantics

The Normativity Requirement demands that inner structures be capable of misrepresenting because the beliefs of which they are constituents can be false. However, even if we ignore the problems with the individual causal theories of content,[9] and allow, for the moment, that a brain state can mean *cow* (or *property of being a cow*), that brain state would not have judgeable content (since it would not be true or false of anything) and, therefore, would not represent or misrepresent anything. The Normativity Requirement remains unsatisfied. In the previous section I suggested that a means to satisfying the requirement might be to marry a Representational Theory of Mind with the Language of Thought hypothesis and suppose that brain states, of the sort the causal theorists try to endow with content, can combine to form structures with propositional content. This option is open to those, like Fodor, who argue for a representational system with a combinatorial syntax and semantics.

However, there are difficulties with this strategy when we recall that what is being questioned is whether it makes sense to posit an inner representational system, for a representational system must, one would think, contain items capable of misrepresenting. The strategy seems to rely upon the assumption that the brain is a representational system prior to showing that it will yield structures with representational content; that is, structures with the capacity to misrepresent. But perhaps the circularity of such a rationale need only

(interpreted) content but he can do so only by jettisoning Realism altogether.

[9] I referred to criticisms of the causal/teleological theory above but should also add Fodor, 1987, pp.104-106 which pursues a different line of objection. For criticism of Dretske's version of the causal theory see Fodor, 1984, pp.40ff and 1990, pp.59ff. Fodor's Asymmetric Dependence theory is criticised by Boghossian, (1991) where he argues that natural kind expressions, though susceptible to causal theories of reference, cannot be naturalized via causal (informational) theories of mental content because of the verificationism implicit in the claim that the meaning of these expressions can be made determinate by such theories. In this, and another article (1989) Boghossian turns Fodor's exploitation of the holism of belief-fixation thesis (to explain why the content of a 'cow' token is not a disjunction of proximal stimuli) against him by pointing out that both a potentially infinite, and non-naturalizable (within the causal theory), set of beliefs must be excluded in specifying the content-fixing relation.

be apparent if we substitute 'symbolic system' for 'representational system' since the reasoning could then be as follows: Brain state tokens can be denoting expressions ('cow' tokens, for example) and, although they lack representational (propositional) content, they can be understood as symbols which can combine with other symbols to form structures with such content. After all, if natural language symbols (like the orthographical/phonological sequence #c^o^w#) can be combined, according to a grammar, to form sentences with propositional content, then surely Language of Thought symbols can do the same.

In order to see what is wrong with the reasoning, we need to make explicit two assumptions it rests upon. These are;

A) that there are internal structural elements, which I have been calling 'brain states' but which the cognitivist would call 'symbols', and these elements have a meaning which they convey independently of any context in which they occur; and

B) that a structure consisting of symbolic elements, combined according to rules, has a meaning which is a function of the meaning of the individual elements.

A) is assumed by Fodor, in conjunction with Lepore (1991), McLaughlin (1990), and Pylyshyn (1988), and would need to be assumed by anyone who accepts their arguments, firstly, that since the semantics of natural language is derived from the intrinsic intentionality of internal states, the systematicity and productivity of natural language must be mirrored in the language of thought and, secondly, that this systematicity and productivity can be explained only by recourse to the assumption of context-independent symbols. As we saw, Fodor, McLaughlin and Pylyshyn put these arguments to use in their attempts to discredit Connectionists models of cognition (see chapter 3, section 2). We also found a commitment to context-independent meaning elements in Fodor's account of natural language acquisition, in terms of language of thought-to-natural language predicate mapping (see Fodor, 1975, pp.79-82 and chapter 1, section 3 b)), where the predicates of both natural language and the language of thought have their meaning independently of the context in which they are employed. Here the construction of hypotheses about the meaning (the truth conditions) of natural language predicates may provide propositional contexts but the mapping is between predicates, *qua* elements of propositions, rather than propositions.

The assumption of A) is often a consequence of assuming B) as Fodor and Lepore acknowledge (Fodor and Lepore, 1991, p.147). However, A) is not a logical consequence of B) even if we remove the psychologism of A) and take the elements referred to there, not as internal symbols but, as words of natural language. We can consistently hold that the meaning of a sentence of natural language is indeed a function of the parts, the words, of which it is composed but reason that those parts do not have a detachable, or context independent meaning because they do not acquire meaning until they are used within a sentence. This is the reasoning behind Frege's principle that 'it is only in the context of a sentence that a word has a meaning' (Frege, 1884, p.x and §§60,62,&106).[10] It may be tempting to claim that because a sentence is formed by combining words (orthographical/phonological sequences), the meaning of the sentence must be formed by combining the context-independent meanings of the words, but, as I will argue shortly, the claim loses credibility as soon as we try to assign such meanings to words.

The context principle is the second of three proposed by Frege in his *Grundlagen der Arithmetik*, the first being that 'There must be a sharp separation of the psychological from the logical, the subjective from the objective' and the third being that 'The distinction between concept and object must be kept in mind' (op.cit., p.x). He indicates the interrelation of the first and second principles by saying of the second that if 'it is not

[10] Assumption A), in its psychologistic form, would be anathema to Frege but Micheal Dummett appears to believe he would assent to it when it is freed from internalist trappings (Dummett,1973, pp.4&5). However, for evidence that Frege did not think that there are, what Ryle calls, 'detachable sense atoms' see James Conant, 1998, pp.231-235.

observed, then one is almost forced to take as the meaning of words mental images or acts of an individual mind, and thereby to offend against the first as well' (ibid.).[11]

In cognitive science the Fregean template for a propositional calculus is often visible as a basis for accounting for inferential relations between the propositional contents of thoughts. However, by interpreting Frege's 'thoughts' as psychological, rather than logical, constructs, cognitive science betrays his first principle. When the interpretation is conjoined with the postulation of internal context independent meaning elements all three principles are betrayed; for if the words of 'Mentalese' have detachable meanings then not only is the second principle contravened, but the third is also. I will explain both why this is so, and why, when Frege's third principle is understood as a grammatical observation, the contravention cannot be allowed. But, before doing so, I will argue that there are no suitable candidates for the meaning of a context independent internal symbol.

b) Context-Independent Meaning

On the causal theory, it is the nomic relation between cows and brain states of a certain type that make tokens of that type mean *cow*, and mean *cow* independently of the context in which they occur. Disjunctive meanings are avoided by adding constraints on the type of nomic relation in which a token actually means its cause, such as the Asymmetric Dependence condition, so that the tokens produced by the appropriate type of relation can be said to mean, or carry information only about one type of cause. But what is the information being carried here? The English word 'cow' does not carry information in isolation from a context, that is, it is not informative until it is used to say something about a cow, or to say

[11] Frege's antipathy towards psychologism is understandable given that his concern in the *Grundlagen* was, broadly, to show that the proofs of arithmetical propositions could be grounded on purely logical principles (which he took to be objective). If the limits of logical validity are understood as contingent upon human psychology, or if truth in logic is contingent upon the historical development of human ideas, then, for Frege, the grounding of arithmetical proof in logic will be a worthless enterprise.

of something that it is a cow, so why should the Mentalese equivalent be thought to be informative?

Note that if the informational theorist changes tack, by urging that a 'cow' token does not carry the information *cow* but something like *a cow is present* (or, if the theorist takes the Fodorian lead, *the property of being a cow is instantiated*), then the theorist will have to forego the combinatorial story. This is because, on the combinatorial story, the relevant brain state representing this content would be complex rather than atomic, which means we will require an account of the content of the meaning atoms forming the complex, particularly the content of the atom which means *cow*. If the account is to be one furnished by a causal theory then we arrive back at the original question of how that atom can have a context independent meaning. *Its* meaning cannot be the informational content *a cow is present* on pain of vicious regress. If the combinatorial story is jettisoned but the representational account is retained (as some Connectionists might have it) then the 'cow' token will cease to be a meaning *atom* and will have the content *a cow is present* instead. Of course, there are an indefinite number of beliefs one can have about cows, an indefinite number of which do not entail any claim about the presence of a cow. The result will be a proliferation of representations, one for each distinct content, and an account of how each content is fixed will be owing. In effect, an explanation of the productivity and systematicity of language will have been lost, and this will be a worrying omission from any theory which recommends itself as an explanation of linguistic competence and inferential reasoning. It is considerations like these that give the insistence on a combinatorial representational system its appeal.

Another response to the challenge of giving an internal element a context independent meaning might be to treat brain states as symbols with extensions, that is, to treat them in the way that many philosophers treat natural language concept, or predicative, expressions (or terms). These expressions, in having an extension, are thought to pick out a class of objects thereby giving them a (referential) content. This response comports with Fodor's (1975) account of natural language acquisition in which hypotheses are formed, in the language of thought, as to which natural language predicates map onto existing Mentalese predicates.[12] However, the treatment of natural language predicative terms as picking out extensions does not support the view that such terms have a context independent meaning because a predicate picks out an extension only when it is used to do so, and that requires that it have a context of use; a context in which it is used as a predicate. We have not done anything meaningful by uttering, or writing down, a predicative expression unless it is being used to say something about what is referred to by an object expression. Indeed, what is uttered or written is not even a predicate until it is used as one. The situation is, clearly, forlorn in the case of the brain state which is purported to instantiate a predicate, for a brain state is not used to say anything.

Further, if, as Putnam tells us 'the *extension* of a term, in customary logical parlance, is simply the set of things the term is true of'(Putnam, 1975, p.4), then 'customary logical parlance' is misleading because a term, on its own, is not *true* of anything. The term 'cow' is not true or false until it is applied to an object by being given a propositional context. The contention that 'cow' is true of all and only cows obfuscates the grammatical observation (on the predicative use of the expression 'cow') which is that a sentence of the form '*x* is a cow' is uttered truthfully (as opposed, not only, to falsely but also to metaphorically, or ironically, for instance) only when it occurs in a context where *x* is used to refer to something which is a cow. In such contexts, therefore, 'is a cow' is predicated only of things

[12] We might observe that, there seems to be a tension between this account of language acquisition and the causal semantics of Fodor's later work. The latter provides an explanation of how internal tokens acquire meaning via their causal relations, but the former presupposes that these tokens have meaning from birth, for natural language predicates are learned by mapping them onto predicates of Mentalese. There could be truth rules guiding this mapping (of the form; [Py] is true iff *Gx*) only if the Mentalese items already had meaning, because the rules establish that these items have the same *meaning* as the items of the natural language.

which are cows, but 'is a cow', when it is not predicated of anything, is neither true nor false. It is neither true nor false it because has not been given meaning.[13]

It is interesting that Fodor moves from talk of extensions, when accounting for the meaning (and, in particular, the reference) of internal states (Fodor, 1975, p.59, for example) to an account phrased in terms of properties. Given his adoption of informational semantics, this is well-advised because, as Hartry Field observes (Field, 1978, p.45), a causal contact with cows could not help in explaining why a 'cow' symbol has in its extension those cows which are not in causal contact.[14] Fodor's causal theory posits nomic relations between 'cow' symbols and the property of being a cow, and this is why he is able to avoid the problem of explaining the meaning of terms lacking an extension (see Fodor, 1990, pp.100&101). But does this advertence to properties help to explain the context independent meaning of internal states? Surely not for, on the one hand, *the property of being a cow* cannot be the meaning of 'cow'. A cow can be said to possess a property but not a meaning.[15] On the other hand, 'cow' cannot have meaning *because* it picks out the property of being a cow, for picking out a property, like picking out an extension (if we accept that there are extensions) requires an appropriate context.

A fundamental problem with taking the internal symbol 'cow' to be a *concept* expression, which refers to an extension or to a property, is that it requires that we see this reference as an intrinsic, rather than relational, property of the symbol. This is not the case in natural language where the same sign (or orthographical/phonological sequence) can symbolize either as a concept or an object expression, and which type of symbol it is will be

[13] And, we might also note that the open sentence 'x is a cow' is not true or false either.

[14] Fodor might be able to avoid Field's objection by appeal to counterfactuals, but not for long. Thus, if an object with the causal power to produce 'cow' tokens (such as a cow encountered at a future time, or a horse on a dark night) was to come into contact with a subject, then the object would cause a 'cow' token. The extension of 'cow' could be narrowed through further use of counterfactuals, to exclude non-cows, by adding the Asymmetric Dependence condition. However, it becomes unclear why 'cow' does not mean whatever has the causal powers to produce 'cow' tokens rather than whatever is a cow unless the Asymmetric Dependence condition is phrased by appeal to something which enables us to distinguish cows from non-cows, in other words, by appeal to the property of being a cow.

[15] Note that the de dicto option of 'the property of being a cow' cannot give the meaning of 'cow' because it is the relation between a property, rather than a sentence, and the token that is supposed to be the one that fixes the token's content.

dependent upon its relation to the context in which it occurs. At this point we can return to my contentions, firstly, that the postulation of context independent meaning elements contravenes Frege's third principle, that a sharp distinction must be drawn between concept and object and, secondly, that, when understood as a point of grammar, the principle should be upheld. Frege's worry, in 'Über Begriff und Gegenstand' ('On Concept and Object'), was that the conflation of concept and object would amount to an identification of the concept expression with an object expression (what he calls a 'proper name'). However, since 'the behaviour of a concept is essentially predicative' and 'a proper name can never be a predicative expression, though it can be part of one', the identification cannot be licit (Frege, 1892, pp.200&201).

Now, the postulation of context independent meaning for a word requires that the meaning be unitary, for if the word had multiple meanings then an explanation of which meaning a word had on a particular occasion, or tokening, would require appeal to its context, which belies the postulate. When we recognise that the word 'cow' can be both an object expression, as in 'The cow', and a concept expression, as in 'A cow' (cf. ibid., pp.195&196) the implication is that the context independent meaning of an internal 'cow' symbol must be an amalgam of both expressions. That is, the 'cow' symbol must be both an object and a concept expression; hence the contravention of Frege's principle. That Frege is correct in rejecting the amalgamation is indicated by the fact that 'cow' does not have the same meaning in sentences like 'The cow is calving' and 'Daisy is a cow', for the first sentence will say something about a particular cow while the second says of something that it is a cow. If, as Fodor thinks, 'cow' denotes the *property of being a cow*, then, since 'cow' and 'the property of being a cow' are co-referring, we can substitute the second expression for the first to render 'The cow is calving' as 'The property of being a cow is calving' and this is clearly an incorrect rendition. Furthermore, even when 'property of being a cow' can be used in an object expression, as in 'The concept *property of being a cow* is a first order concept' the expression 'property of being a cow' has to be converted into an object expression by being given the subject position in the sentence and by being modified by a prefix (the words 'The concept'), italics, or some other such device (see ibid., p.197). The position and the modification show how the expression is being used, thereby changing the meaning by changing the propositional, and sentential, context.

At this point we might consider whether Fodor's theory of content has any resource with which to meet the objection. I believe there are two sorts of reply Fodor might make. The first is to try to show how the two types of expression can be accounted for by the nomological relations between symbols and their causes, and the second is to argue that in the internal symbol economy there need only be trade in concepts.

In *Psychosemantics* when outlining the basic form of causal theories, to which he will add his asymmetric dependence condition, Fodor writes;

> 'Let's start with the most rudimentary sort of example: the case where a predicative expression ('horse', as it might be) is said of, or thought of, an object of predication (a horse, as it might be). Let the Crude Causal Theory of Content be the following: In such cases the symbol tokenings denote their causes, and the symbol types express the property whose instantiations reliably cause their tokenings. So, in the paradigm case, my utterance of 'horse' says *of* a horse that it *is* one.' (Fodor, 1987, p.99)[16]

Using a causal theory of content, it would seem, we can have the same symbol acting as both an object and a concept expression. The symbol type will mean *property of being a horse* and tokenings of the type will mean *instantiation of the property of being a horse.* This is compatible with the Asymmetric Dependence Theory in which it is the nomic relation between the property of being a horse (rather than the property of being a non-

[16] Although Fodor talks of 'utterances' here and, therefore, seems to be concerned only with the semantics of natural language, he argues that 'it is mental representations, and not the formulas of any natural language, that are the natural candidates for being the primitive bearers of semantic properties' (ibid., p.100). The phrase 'utterance of' in the last sentence of the quotation in the text, therefore, can best be read as 'thinking'.

horse) and the property of being a cause of 'horse' tokenings, which allows us to say, on the one hand, that the 'horse' symbol *type* means the *property of being a horse*; that is, it is a type of concept expression picking out a property. Since, on the other hand, an individual 'horse' *token* is caused by a particular instantiation of the property of being a horse it seems reasonable to maintain that that gives us the token's meaning as an object expression.

The unacceptability of this line of thinking soon becomes apparent when considering the last sentence quoted. Here a 'horse' token is supposed to predicate of something that it is a horse, and we have just seen that a 'horse' token means *instantiation of the property of being a horse*. So, since 'horse' and 'instantiation of the property of being a horse' refer to the same thing, it should be permissible to rephrase the (internal) sentence 'Bucephalus is a horse' as 'Bucephalus is an instantiation of the property of being a horse'. However, as the token denotes its cause, which would be a particular instantiation of the property, what we find is that the 'is' of the original sentence cannot be the copula for it connects two objects, *viz.* the object denoted by 'Bucephalus' and the property instantiation denoted by 'horse'. That is, the original sentence turns out to be an identity statement, whether or not it was meant to be one, and we are without an account of how 'horse' can be a predicative expression.

The only way to avoid this consequence, it would seem, is to take the expression 'instantiation of the property of being a horse' as expressing a property itself. It would express the *property of being the instantiation of the property of being a horse* which would seem to have the virtue of allowing the expression to pick out an object as well as a property. Now, if this was thought to be the content of a 'horse' token (as it must be if the token is to be predicative) then it would follow that the nomic relation which provides this content is between the property of being the instantiation of the property of being a horse and the property of being a cause of 'horse' tokenings. Since the instantiation of the property of being a horse just is a particular horse, let us call it 'Dobbin', it follows that the nomic relation is between the property of being Dobbin and the property of being a cause of 'horse' tokenings. The intolerable consequence is that all tokenings of 'horse' which can occur as predicative expressions in sentences say of the sentential subject that it has the property of being Dobbin.

The problem cannot be avoided by changing the first definite article, in 'the instantiation of the property of being a horse', for the indefinite article 'an', for although this may seem to allow that a 'horse' token can mean the *property of being any horse* and, therefore, provide the requisite generality, two considerations arise. Firstly, a singular relation (instantiating the nomic relation) cannot be between the property of being a cause of 'horse' tokenings and the property of being *any* horse, for a singular relation will have to be between determinate property instantiations. Thus, it will not be any horse that causes the tokening but a particular one. Secondly, if the 'horse' token occurred as an object expression, as in the sentence 'The horse is a mare', the sentence would almost always be false because it would say 'Any instantiation of the property of being a horse is a mare'. The sentence would be true only in contexts when all the relevant animals were female and so could not be true in many contexts where 'The horse is a mare' is true.

The conclusion is that whichever way the content of 'horse', or any other symbols, is fixed they cannot occur as both concept and object expressions. Fodor's assignment of predicative content to symbol *types* is of no help because, although this would free tokens from predicative duties, it is always tokens of types that occur in sentences. Furthermore, when we think of types as classes we can see that it makes no sense to say that our exemplar sentences contains classes of symbols. We might also note that the problem is not circumvented by positing separate types for symbolizing objects on the one hand, and properties on the other. We might be able to make sense of the causal account of the content of the latter (if we ignore the problems involved in supposing they pick out properties independently of context), since the content of symbol types is based on nomic relations between properties. However, in the case of object expressions, the relevant properties will be the property of being causes of tokenings and the property of being the (particular) object doing the causing (for it is a particular object which causes a given token) from which it follows that a different symbol type will be required for each object denoted. The profluence of symbols would flood the representational economy and wash away any explanatory value.

The second strategy Fodor could adopt in dealing with the problem might be to maintain that the meanings of the internal symbol elements just are concepts, and are, therefore, unitary because it is unnecessary for any symbols to be object expressions. The rationale would be that we already have a model for such a representational system in the form of predicate calculus. For, in the predicate calculus, we can symbolize any closed sentence by replacing singular terms, or object expressions, with bound variables and the expressions predicated of them. Thus the sentence 'The horse is a mare', in which 'horse' seems to be an object expression, can be symbolized as:

1) $(\exists x)\ [Hx\ .\ (y)\ (Hy \supset x = y)\ .\ Mx]$

The formula can be expressed as 'There is one, and only one, thing such that it is a horse, and that thing is a mare'. After analysis, then, 'horse' occurs as a concept expression, where it is used in assigning an object to the extension of the concept *horse*, rather than as an expression for an object of predication. The sentence 'Bucephalus is a horse', in which 'horse' is clearly used predicatively, is analysed simply as:

2) $\mathrm{H}b$

So, both sentences can be symbolized in such a way that 'horse' is to be used as a concept, or predicative, expression thereby supporting the contention that if our internal language was modelled upon the predicate calculus, all we would require to understand sentences of the natural language are representations which function as concept expressions (perhaps together with some proper names). The implication would seem to be that the meanings of the representations and, hence, the words of natural language, are concepts. Fodor recently indicated that this is his belief when, in introducing his discussion of concepts, he explained that he 'proposed to move back and forth freely between concepts and word meanings' because 'for the purpose of ... investigation word meanings just are concepts' (Fodor, 1998, p.2). He also writes that 'for present purposes, it will do to think of thoughts as mental representations analogous to closed sentences, and concepts as mental representations analogous to the corresponding open ones' (ibid., p.25). The juxtaposition of the two expeditious statements gives the impression that Fodor sees concepts as playing the role of both the meaning *and* symbolic constituents of language of thought formulæ. Thus, concepts will be the symbolic elements (of the form; Fx) constituent of internal formulæ, or thoughts, of the form; ($\exists x$) (Fx). Indeed, since Fodor says that 'concepts are constituents of mental states' and 'the concept ANIMAL is a constituent of the belief that *cats are animals*' (ibid., p.6) the impression seems to be correct.

The problem arising from the argument is that if understanding a sentence is a process of internal analysis yielding structures like 1) and 2), then we could never understand sentences, like the two just quoted, which have concepts as their subject matter. Understanding such sentences would require not only that we quantify over the conceptual constituents of internal formulæ, but also that we bring those constituents under concepts.

In the case of quantification, the problem is that a concept symbol cannot be a bound variable because the bound variable is a complete symbol. That is, there is not a place in the symbol for another variable (or a constant, in the case of a proper name) to occupy and there should be such a place if the symbol is a concept. Thus, our first sentence 'Concepts are constituents of mental states' cannot be rendered as;

3) (x) $(Fx \supset Gx)$

This says of all *x*s that if they are concepts then they are constituents of mental states but, for the antecedent to be satisfied, all *x*s will have to be concepts. However, since concepts are to be incomplete symbols, they cannot be *x*s. The problem is circumvented only by introducing second order quantifiers which range over concepts rather than the complete entities falling under them. That a quantifier is of the second order must be shown by symbolizing concepts differently from objects. The only way to make plain that we are symbolizing a concept is to give the symbol an argument place, thus: *Fx*. This indicates that the symbol means a concept because there is a position, marked by the variable, in which to place an expression standing for an object falling under the concept. (If we are to do away with object expressions then we will need to retain the incompleteness of the predicative expression. Only in this way will we afford the analysis of object expressions (other than proper names) into predicates (names of concepts) and the bound variables which complete them.) As it stands, however, the symbol '*Fx*' is incomplete, for the variable, being unbound, is to symbolize any object, which means that no particular object is being said to fall under the concept. This is as it should be since if the variable was bound then *Fx* would express a proposition rather than a concept. However, since the variable can serve only to mark, and not fill, the argument place we should reproduce the symbol as *F*( ) or *F*... . By doing so we make explicit its incompleteness as a symbol and, therefore, its incompleteness as a concept. We cannot present the same concept in two different ways because, since this symbol is to have a unitary meaning, it must be structured such that it can fulfil the same role—for we must remember that these symbols are to have a causal role within the organism's representational system (or language of thought), and the 'shape' of the symbol will be a determinant of its role.[17]

For similar reasons the internal equivalent of the sentence 'The concept ANIMAL is a constituent of the belief that *cats are animals*' cannot be:

4) $(\exists x)\ [Fx\ .\ (y)\ (Fy \supset x = y)\ .\ Gx]$

Again, this is unsatisfactory because we are not quantifying over objects; that is, since first order quantification cannot be used for concepts we cannot be quantifying over concepts in 4). The embedded identity relation, which contributes to rendering the bound variable in 1) as a definite object, cannot be used for concepts expressions which, unless predicated of a bound variable (in which case they are propositions), are incomplete symbols and, therefore, unsuitable for inclusion in identity relations. The variables in 4) are complete expressions and that is what predicate variables cannot be in the language of thought. Thus, quantifying over predicate variables in the internal language requires that we attach sense to expressions of the form 'For all ... is an *F*' and 'There is some ... is an *F*', and clearly we do not. Even if we allow that being well-formed is sufficient for having a sense (which we should not) then these expressions are clearly ill-formed and, therefore, nonsense. Furthermore, the sentence we tried to formalize in 4) cannot be rendered as 'There is some ... is an F such that *it* is a G', for we do not have anything for 'it' to refer to.

The sentences 'Concepts are constituents of mental states' and 'The concept ANIMAL is a constituent of the belief that *cats are animals*' presumably predicate of concepts—all concepts, in the first place, and the concept ANIMAL, in the second—that they fall within the extensions of other concepts. In the symbolism of the internal language this amounts to the argument place of one concept being filled by another concept. The problem is that, as

[17] In attempting to explain the formal/syntactic properties of symbols Fodor tells us that 'formal operations apply in terms of the, as it were, shapes of the objects in their domains' (Fodor, 1980, p.227). Thus, 'the syntax of a symbol might determine the causes and effects of its tokenings in much the way the geometry of a key determines which locks it will open' (Fodor, 1985, p.22).

an incomplete expression, the symbol *F*( ) cannot fill the argument place of any other such expression to express a proposition. The sentence 'The concept ANIMAL is a constituent of the belief that *cats are animals*' uses a complete symbol to express the concept and, for this reason, does not properly express it, for a concept is incomplete. The internal symbol, because its meaning is unitary, must be the same whether it is in the argument place of another concept or not. Thus, the correct rendition of the sentence as it is represented in the internal language would be '... is an animal is a constituent of the belief that *cats are animals*' but, again, this is not a sentence.

Note that the rendition would not be '"...is an animal" is a constituent of the belief that *cats are animals*' which is, at least *prima facie*, well-formed since '... is an animal' is here behaving as a *name* for a string of symbols. As we have seen, Fodor seems to think that the symbolic constituents of mental states are, themselves, concepts so, in the language of thought, the symbol '... is an animal' is not something which completes a concept, as an object does, but is a concept itself. If, by putting inverted commas, or the internal equivalent, around the symbol it is made into something other than a concept then, whatever role it has subsequently, it is not the role of a concept. In other words, if the concept, *qua* symbol, is something having a causal role by virtue of its 'shape' or form, and the concept, *qua* meaning, is incomplete, as it must be if it is to function predicatively, then to complete the symbol so that it has the role of a name is to complete the meaning. So, when such symbols are in the subject place of a sentence they are either not concepts or, if they are, they are not in sentences after all. The only way round this would be to give two separate meanings to a concept symbol and allow that which meaning was being conveyed by the symbol is to be determined by the context in which it occurs. However, Fodor's meaning atomism cannot allow this.

The paradoxical situation is that if the two sentences we have tried to formalize are true then they are nonsensical. In other words, if the internal language is constructed solely from concept expressions and object variables then it cannot represent concepts as logical subjects. Since, on Fodor's hypothesis, understanding a proposition, expressed by a sentence, just is to have a certain relation to a representation of that sentence it follows that we cannot understand propositions about concepts. It seems fair to say, then, that we cannot make sense of the claim that concepts are the constituents of mental states, such as understanding that *p*, and, hence that concepts are the context independent meaning elements of the language of thought.

Although the arguments have been directed specifically at the conjunction of the Asymmetric Dependence Theory of mental content and the Language of Thought hypothesis, I think it extends to any causal theory so conjoined. Furthermore, it is difficult to see how a non-causal theory could retain the atomism many assume is required to deal with the problem of linguistic creativity; the problem for which the combinatorial language of thought is invoked as a solution. For example, Conceptual Role Semantics, if devoid of content-fixing relations between symbols and causes, seems to be incompatible with atomism because of its holistic implications—for it will not allow a symbol to having a meaning independently of relations to other symbols.

That is not to say that a meaning holism is incompatible with attempts to solve the problem. Donald Davidson, for example, relies on such a holism in his attempt. Neither is it to say that compositionality entails meaning atomism, for Davidson uses the notion of the satisfaction of predicates to make open sentences (of the form *Fx*) the units of meaning (see Davidson, 1984). But Davidson, unlike proponents of representationalism, does not argue for the priority of mental content in explaining the semantics of natural language. In saying that the latter is provided for by the former, representationalists, like Fodor, cannot *assume* the semantics of a metalanguage of internal representations in which to give interpretations of natural language sentences, for their thesis of explanatory priority shackles them to the burden of providing an account of how internal representations have original meaning. And, as we have seen, they, unlike Davidson, cannot rely on interpretation as a means to providing that account, which means that representational content must be built from symbolizing elements. If these elements are to have context-independent meanings then it is causal theories that offer an explanatory route to these. If they are not, then it is difficult to see how they can contribute to the construction of internal structures with representational content.

I have been arguing that there is no such thing as the context-independent meaning of an internal state or, in other words, there are no such things as internal meaning atoms.[18] Assumption A) is, therefore, false insofar as it attributes context-independent meaning to internal symbols. If it is false in this respect then assumption B) is false as well when it is claimed that the structures with meaning are internal. When the structures in question are natural language sentences it is not so much false as misleading because to combine symbols grammatically is not sufficient for the combination to have meaning. One might, for example, construct a sentence by combining a noun and verb phrase and, provided one pays attention to matters such as the tensing of verbs and the numbering of nouns, the sentence will be grammatically correct. But, its having a determinable meaning is a matter of its actually saying something, and this requires a situation, or context, in which

[18] The debate has been about the context independent meaning of concrete nouns such as 'cow' or 'horse'. The theorist opting for a combinatorial semantics of language will need to defend the claim that other parts of speech, such as abstract nouns, adjectives, prepositions, and pronouns, have a context independent meaning. For Fodor, the meaning of an adjective like 'virtuous' is 'determined by the nomic relation between the property of being a cause of tokens of that word and the property of being virtuous'. That is, 'It isn't interestingly different from the semantics of "horse"' (Fodor, 1990, p.111). Ran Lahav (1989) presents objections to the assumption of the context independent meaning of adjectives based on the observation that their meaning is noun-dependent. He points out the noun-dependency of the meaning of prepositions and many verbs whilst doing so.

something is being said by uttering or writing down the sentence, and the mere constructing of a grammatical sentence does not constitute such a situation.[19]

Of course, when a sentence does say something, what it says will be determined, in part, by the words of which it consists. Thus, the sentence 'A cow is in the field' is meaningful in an appropriate context, such as when it is a reply to the question 'Are there any animals in the field?', and the use of the word 'cow' is integral to the meaning of the sentence in this context—if 'horse' was used then the speaker would be saying something different—but how the word 'cow' is being used is dependent upon the context of the sentence, and that it *is* a word is dependent upon its being used, in a sentence, in an appropriate context. What applies to words, here, applies to symbols in general. The meaning we might give to a symbol depends on how it is used, and this context of use not only shows what it means (what it symbolizes, refers to, requires, stands for, and so on) but also that it *is* a symbol.

However, there is still a sense in which the parts of a sentence can be called 'symbols' independently of a context of use. Orthographic/phonological sequences can be called 'symbols' because they are recognisable as written or spoken items which can be *used* as symbols, as words, in a sentence. This, however, does not imply that they symbolize anything in isolation, that they, so to speak, carry their meanings with them. One might, whilst indicating the position of the fire exits in a certain building, use a tea caddie to symbolize, or represent, the building (designating one side of the caddie as the building's front). But this does not mean that the tea caddie *always* symbolizes the building. Words, unlike tea caddies, are used, primarily, to say things, so even when they occur outside a context in which they are used *as* symbols we can associate them with contexts in which

[19] I take this to be a generalization of one of Wittgenstein's insights in On Certainty (see Wittgenstein, 1969, §§347-352 for example). I would suggest that the temptation to suppose that a given sentence can be understood independently of a context arises from the conviction that we know how to use the constituent words or that we can easily imagine a context in which they would be used. However, such a conviction does not imply that the sentence has a physiognomic meaning and, indeed, it could not, for the same sentence can be used to say very different things.

they say something, and can provide definitions based on these uses; such provision is the function of dictionaries. But words do not determine their use because of any meaning belonging to them intrinsically; if they did then the same word could not be used as both an object and a concept expression. The word 'cow', for example, can be used to say of something that it is a cow, and to say something about a particular cow. To suppose that 'cow' in the first instance meant the same in the second would entail either that there is no predicative use of the word—for 'Daisy is a cow' would have to be a statement of identity if 'cow' meant 'a particular cow'—or that there is no way of using the word to refer to a particular, so that 'The cow is calving' would have be a statement about the species. That the word does not have a unitary meaning is demonstrated, further, by the fact that 'The cow is a cow' can be an informative utterance—when, for example, what is in question is whether a particular cow is a male or female of the species.

This indicates the need for vigilance when speaking of publicly employed symbols independently of a symbolic context, but it is not vigilance but dismissal which is appropriate when encountering talk of internal symbols. The contexts of use with which we might associate marks and noises are not applicable to brain states and, for this reason, it is nonsense to say these are 'symbols' in any sense. That is to say that internal states not only lack context independent meaning, but also lack appropriate contexts of use in which they could have (as marks and noises do have) a context *dependent* meaning.

c) Conceptual Role Semantics

A response to this might be that Mentalese items are worthy of the epithet 'symbol' because they play the role of meaningful items. That is, they make contributions to the meaning of internal sentences, and to the inferential relations that exist between these sentences, and between these sentences, sensory inputs and behavioural outputs. Such is the claim of Ned Block who, in his 'Advertisement for a Semantics for Psychology' (1986), urges us to accept that 'what makes an expression meaningful is that it has a conceptual role of a certain type' (p.114). Conceptual role, according to Block,

> 'is a matter of the *causal role* of the expression in reasoning and deliberation and, in general, in the way the expression combines and interacts with other expressions so as to mediate between sensory inputs and behavioural outputs' (op.cit., p.93, my italics.),

and 'the conceptual roles of external language are inherited from those of internal language' (ibid., p.129). So, for the advocate of Conceptual (or Functional) Role Semantics, internal items are worthy of the epithets 'symbol' and 'word' because they have meaning by virtue of the role they play. To take an example from Block,

> 'the conceptual role of "and"...derives from such facts as that a commitment to rejecting "*p*" (in the absence of a commitment to accept "*p* and *q*") can lead (in certain circumstances) to a commitment to rejecting "*p* and *q*"' (ibid., p.132).

Whether Conceptual Role Semantics requires, or entails, a meaning atomism of the kind embraced by those, like Fodor, who like their semantics compositional, is not immediately clear. On the one hand, if it is its role in sentential, and supersentential (inferential) contexts that makes an internal item meaningful, then it need not be supposed that the item has a meaning independently of such contexts; it is the contexts which fix the meaning of the item. On the other hand, Block tells us that 'a word's conceptual role is a matter of its contribution to the role of sentences' (ibid., p.93) and this suggests that words bring something to the meaning of the sentences they form. It would seem that the contribution words make to sentences is the semantic value implicit in their conceptual role, for Block tells us that 'CRS explains why words have the conceptual roles they do by appeal to conceptual roles of sentences; thus the semantic values of words are seen to be a matter of their causal properties' (ibid., p.132). Since Block conflates the conceptual and causal role of words and explains a word's semantic value in terms of its causal properties, it seems sensible to assume that insofar as the causal properties, and causal role, of a word is fixed, so too is the meaning of the word. Thus, if those properties are intrinsic, as one might think, rather than relational, so that the possible causal roles of the word are determined by the constitution of its (presumably neural) realization, then this will amount to the supposition that it has a fixed meaning which is independent of context. This reading of Block is secured by his commitment to the view that 'representational states themselves constitute a combinatorial system' (ibid., p.106), a view he adheres to because it seems to be the best way to explain the productivity of thought (p.107).

If, as seems to be the case, Block is viewing the conceptual role, and meaning, of a word as detachable from its role in a given sentence then his account of mental, and linguistic, content will be subject to the objections, raised above, against such a position. His meaning atomism will require that words have meaning in isolation, that words will refer to, or pick out, an object, extension, or property even when they are not used to do so, and that the same word will be an object and concept (a subject and predicate) expression simultaneously, and it will be as unintelligible, because of these requirements, as was Fodor's atomism. However, there is another aspect of Conceptual Role Semantics, one I believe is equally unintelligible, that I wish to draw attention to.

As we have seen, Block claims that the conceptual role of an expression is what gives it meaning.[20] Thus, in assigning a meaning to an expression we are to observe the relations it enters into with other expressions and the results of these relations. For example, in an inference of the form;

[20] In fact, Block advocates what he calls the 'two factor' version of Conceptual Role Semantics 'in which the conceptual role factor is meant to capture the aspect (or determinant) of meaning "inside the head," whereas the other is meant to capture the referential and social dimensions of meaning' (1986, p.101). Block wants to include an external component to meaning, one concerning relations between representations and their referents in the world, in order to avoid the traps set by Putnam and Burge's Externalist arguments, and to get around difficulties regarding indexicals. The component of meaning provided by conceptual role is narrow meaning but this component is supposed to be essential to an understanding a speaker's meaning since it determines a function from context to reference and, therefore, truth conditions. That is, it allows us to map what is in the head onto the world (once we have an adequate theory of reference).

1) $p?q \vdash p$

we can deduce that the symbol '?' means 'and' because, as far as the logical connectives are concerned, it is only in the case of conjunction that the truth of the compound proposition entails the truth of either of its components. On the assumption that 1) is a valid inference, the inferential role of '?' has allowed us to assign it a meaning because only if it means 'and' will 1) be valid. The first point I wish to highlight is that the assignment of meaning to the symbol '?' presupposes a normative framework, a framework of norms, or rules, without which the symbol in question would have no determinable meaning. That is, unless simplification (or 'and' elimination) was an accepted rule of inferential reasoning it would not be possible to derive the meaning of '?' from 1). Conceptual Role Semantics, therefore, presupposes that symbols have a role in a normative framework in order to claim that a symbol's conceptual role will determine its meaning.[21,22]

The second point I wish to draw attention to is that, in claiming that conceptual role is a matter of causal role, the conceptual role semanticist is, in effect, claiming that the normative framework necessary for a symbol's having a meaning can be redescribed in terms of causal relations between symbols. Furthermore, if the conceptual role semanticist is to offer a naturalistic account of mental content then it must be that the normative framework is *explicable* in terms of causal relations. That these are Block's claims is obvious from his self-inflicted interrogation;

[21] The conceptual role of a symbol will, presumably, be aggregative in that it will constitute a totality of actual and potential conceptual relations that the symbol will, or would, enter into. Thus, the symbol '?' will mean 'and' if, in addition to its occurrence in 1), it would also occur in the inference 2) $\sim p \vdash \sim(p ? q)$; and if it would play a role in causing behaviour such as a person's reaching for an umbrella on being told 'It's cold and wet outside'. Of course, such aspects of a symbol's conceptual role will also presuppose a normative framework. In the second example, reaching for the umbrella is the rational thing to do if one knows that it is cold and wet and does not wish to get wet, and if rational action is to be relevant to the conceptual role of an internal symbol, it is because it conforms to norms of practical reasoning.

[22] I have taken a logical connective as my example because it seems the least problematic from the point of view of Conceptual Role Semantics. If, for example, we were to consider 'and' as a sentential connective, then we would have to find some way to distinguish its conceptual role from its inferential role, for, as far as inference is concerned, the sentential connectives 'and', 'but', and 'although', are the same, whereas from the point of view of meaning, they are distinct. If conceptual role, or meaning, is the same as inferential role, therefore, the distinctions cannot be accounted for. Regarding the non-logical vocabulary, Block seems to believe the meaning of its items is also fixed by conceptual role. Thus, part of the meaning of one's internal symbol 'Tiger' will be provided by the fact that one infers from 'Tiger' to 'Dangerous' (see ibid., p.94). This might seem puzzling to those of usused to thinking of inference as being between propositions rather than terms.

> 'How does the brain confer meaning on its representations? Answer: by conferring the right causal roles on the representations. What is it for a person to grasp the meaning of a word? Answer: for a person to grasp the meaning of a word is for the word (or its standard Mentalese associate) to have a certain causal role in his or her brain' (ibid., p.128).

If conferring meaning upon a symbol is a matter of giving it a role, which conforms to norms for the use of a symbol with that meaning in inferential contexts, and the role the symbol is given is a causal one, then it should follow that the causal role will accord with norms for the symbol's use. If this did not follow then there would be no sense in talking about the 'right causal roles', for the use of 'right', as a normative evaluation, implies that the causal role of a symbol is, in fact, a normative role. Furthermore, since grasping the meaning of a symbol surely involves being able to use it correctly, and if, for Block, a symbol is used correctly if it has a certain causal role in the brain, then it follows that correctness of use is something to be understood in terms of the causal relations into which a symbol enters.

In short, the explanation of meaning by appeal to a symbol's causal role is coherent only if that role is indistinguishable from the role a symbol has in a normative framework, for if the causal and normative roles come apart then the symbol's relations will not be inferential, or conceptual, and will not fix its meaning. Since the explanation is to be naturalistic, it is the causal role of the symbol which must explain its meaning. But then normative evaluations of its role, such as correctness and validity, will be determined by its causal relations with other symbols, sensory inputs and behavioural outputs. In the next chapter I will offer arguments to the effect that one cannot explain the normativity of language use and reasoning in terms of the causal properties of a cognitive system, or mechanism. Before doing so I wish to stress the importance of considerations of normativity to the evaluation of the idea of a cognitive science.

5 NATURALIZING NORMATIVITY

The request for a naturalistic explanation of normativity from cognitive science is imperative. Language use, reasoning, and rational action are all phenomena in which performances are evaluated in the light of norms, standards and rules. They are also the explananda of theories of cognitive science. In suggesting that these phenomena are, or are explained by, processes occurring within cognitive subjects the theorists are obliged to demonstrate how these processes and their outcomes are also apt for normative evaluation. Indeed, this is the fully-fledged normativity requirement incumbent upon the cognitivist's theoretical assumptions.

The threats, of vicious circularity and of inducing an infinite regress, against positing internal representations in explanations of cognition, pushes the cognitive theorist towards naturalism to the extent that the explanation of normativity cannot employ cognitive concepts as theoretical primitives since this would presuppose the very phenomena to be explained. The cognitive scientist cannot, for example, explain my ability to understand a sentence in terms of my representing that sentence internally unless the relation of that representation to my behaving *appropriately*, for one who understands the sentence, can be explained without recourse to the concept of understanding or one of its cognates. The explanatory resource to which the cognitivist will make appeal is, as we have noted, the causal power of the representation. Similarly, the ability to speak or write in grammatically *correct* sentences cannot be explained in terms of my knowing the relevant rules of grammar unless my knowing them can be redescribed as my being in a naturalistic relation to the rules such that I will be likely to produce correct sentences. Again, my ability to *correctly* infer from two sentences, expressing the propositions $p \supset q$ and q, respectively, the truth of the sentence expressing q, cannot be accounted for by my employment of the *modus ponens* rule unless this employment can be redescribed as a process the description of which does not use the concept ‘inference’, or any other intentional concept. In both types of case the redescription will be in terms of relations and processes occurring within parts of my central nervous system and, for this reason, though the level of description may be functional or computational, the relations between inputs and outputs of the parts will be relations of cause and effect. Amongst philosophers of cognitive science the parts of the system in which these processes and relations occur are commonly called ‘mechanisms’.

So, for the cognitivist, linguistic competence, reasoning, and action will be explained by appeal to mechanisms. Thus, Fodor’s ‘compilers’ are mechanisms fulfilling a host of functions such as the translation of natural language predicates into their language of thought counterparts—a reversible function responsible for both the learning of the natural language and the production and understanding of natural language sentences—and the translation of sensory stimuli into well-formed language of thought formulæ (see Fodor, 1975, and chapter 1, section 3 b)). Fodor also talks of ‘mechanisms of belief change’, by which he means mechanisms of inference from certain beliefs (like the belief that $p.(p\supset q)$) to others (q) and, more generally, of the mechanisms that implement psychological laws (Fodor, 1987a, p.145). Block assumes that there are ‘mechanisms of language production and language understanding’ (Block, 1986, p.98) and Millikan, to take an example of a representationalist of a somewhat different persuasion, posits mental state producing and using mechanisms, the first of which are required to map mental states onto aspects of the world, and the second of which fix the content of those states according to the proper function of the mechanism (Millikan, 1989).

In what follows I will take a mechanism to be a device for transforming an input into an output in which the transformation is effected by physical processes. In doing so I will use ‘mechanism’ in a wider sense than usual since, according to William Bechtel,

> 'The mechanistic approach to explanation is widely accepted in many disciplines where scientists have tried to explain the behaviour of complex systems by decomposing them into components and then showing how the behavior of the system arises from the behaviour of the components.' (Bechtel, 1987, p.254).

Bechtel makes this observation in order to suggest that Connectionist models are not mechanistic insofar as 'the contributions of the components are minimized and the behavior of the system results more from the interaction of the components than from the behaviour of the components themselves'(ibid.). As we have seen (in chapter 3) the units in a Connectionist network cannot be 'viewed as making a discrete kind of contribution to the behavior of the whole system' and so we need not find components corresponding to Fodor's 'compilers', for example, within the network.[23] However, on my broader understanding of 'mechanism' the whole network can be viewed as a cognitive mechanism since it will transform inputs (such as sightings of cows in fields and hearings of utterances about them) into outputs (utterances and actions) via a physical process, albeit one distributed across the whole network. Now, if the Connectionist system is to model cognition (linguistic competence and inferential reasoning) and explain action, then the relation between inputs and outputs will be such that the latter will have a normative evaluation in the light of the former. Consequently, the Connectionist will need to explain how, for example, the distribution of connection weights and biases within a network will be relevant to the description of its output as according with norms of action, reasoning, and language use.

In general, since cognitive mechanisms will be responsible for explaining phenomena like the production of grammatically correct sentences, appropriate behaviour, and valid inferences, and these phenomena are identified as such because of their conformity to norms or rules, the relations and processes which generate the mechanisms' outputs will need to

[23] We should remember that Clark, an advocate of Connectionism, sees the need to posit 'internal mechanisms' for the storage and retrieval of knowledge in order to distinguish a 'True Believer' from a 'Giant Lookup Table'.

conform to norms and rules also. Thus philosophers of cognitive science are required to provide an account of what this conformity consists in, that is, of why the output of cognitive mechanisms can, and should, be evaluated normatively. I have contended that this is required for us to accept the attribution of a meaning to an internal structure, on the grounds of that structure's purported causal role but, in fact, it is difficult to overemphasize the general importance, for cognitive science, of providing a cogent explanation of the normativity of language understanding, and use, in terms of causal processes. Thinking, talking, and agency in general, are phenomena for which judgements and explanations of the correctness, or otherwise, of what is thought, said, and done are integral to their being the phenomena they are. Since cognitive science seeks a generalized mechanistic account of them it must give an account which allows for this normativity. Without such an account it cannot be explaining language and cognition.

In the next chapter I will argue that the form of the account the cognitivist must give is not one that can be given intelligibly.

CHAPTER 6

NORMS AND CAUSAL PROCESSES

The course of the discussion in the last two chapters has been convoluted so I will try to give a rough outline of the reasoning which concludes that the naturalization of normativity should be demanded of cognitive science. We began the critique by noting that the Representational Theory of Mind promised to support propositional attitude Realism by impregnating internal states and processes with intentional content. On the introduction of the infinite regress argument, however, we found that in justifying the attribution of content to internal states the cognitivist could not avail herself of intentional concepts, such as 'understanding' and 'interpreting'. The attribution, therefore, would have to be justified by appeal to natural, or causal, meaning.

I argued that causal theories of content fail to provide a viable account of representational content unless supplemented by an assumption as to the combinatorial structure of internal representations. But we found that the positing of a combinatorial system did not afford the construction of representational contents because the notion of a detachable meaning element could not be sustained. This left the option of taking internal structures to be the vehicles of representational content on the grounds that the causal and conceptual role of these structures would display an isomorphism and, thereby, warrant our identifying them as sentences. However, specification of the contents of internal structures, by appeal to conceptual role, presupposes a normative framework within which to determine these contents. Since the conceptual role of the structures is, supposedly, explicable in terms of their causal relations and properties, it follows that the framework of content-fixing norms *should* have a naturalistic description which replaces normative notions (such as logical consequence) with naturalistic ones (such as causal consequence). In effect, what is required is an explanation of how causally driven processing can be normatively evaluated.

This normativity requirement must be fulfilled not just because without it there can be no conflation of causal and logical role—that is, no conflation of the causal role of internal structures with the legitimate *use* of symbols—but also because normative phenomena, such as reasoning and language use, are the central explananda of cognitive science. These phenomena are treated as identifiable with, or the products of, mechanistic processes so, again, an account is owed of why we should concede that normative evaluations should be given to the outputs of mechanisms on the strength of the causal processing occurring within. In what follows I will be concerned with the general explanatory enterprise of giving a mechanistic account of normative phenomena and will return to the issue of representational content in chapter 7.

1 MECHANISMS AND COMPLIANCE WITH NORMS

The explanation of cognition in terms of the internal processing of symbols seems to bring with it the bonus of an explanation of linguistic competence; of what it is to be able to speak and understand a natural language. Of course, in doing so it increases the need to render causal and normative explanation compatible. Uses of words are evaluated according to norms and rules in innumerable ways. What one says can be grammatical or ungrammatical, appropriate or inappropriate, well or poorly reasoned, justified or unjustified, accurate or inaccurate, and so on. The norms, standards, and rules which are appealed to in judging a use as correct can have precise or imprecise formulations and can be incontrovertible or contentious. Grammatical rules and rules of logical inference are the ones most relevant to our discussion,[1] but informal norms of reasoning, norms guiding manners, and specific norms of expression found within various areas of discourse (such as scientific research, law, and even philosophy) all provide grounds for the evaluation of forms of expression as correct or incorrect. When a use of words is judged to be correct it is because it conforms to, accords or complies with, one or more norms, standards, or rules.[2]

Depending on the nature of the theory which posits them, cognitive mechanisms will either engage with language at their peripheries (where the inputs and outputs will have a linguistic form—monadic functionalism and some types of Connectionism can be viewed in this way), or will undergo processes ranging over states with a linguistic structure (which could be purely syntactic, as Stich would have had it, or both syntactic and semantic, as Fodor would maintain). In either case, uses of words, in utterances or inscriptions, result from mechanistic processes occurring internally and, in the latter case, internal 'words' have a use during processing where their syntactic properties will determine the process outcome. At any rate, this is generally the picture of language use that cognitive science offers.

The production of linguistic expressions by cognitive mechanisms will either comply or fail to comply with norms of grammar and inferential reasoning but, since linguistic competence, rather than incompetence, is the salient explanatory task of cognitivism, it is the nature of the *compliance* of the mechanism's output with rules that we need to examine.

[1] This is because we are examining the supposition that language use and cognition are explicable mechanically, and formal rules, since they relate to symbol manipulation, appear to be amenable to mechanical application. In one sense this is right for we certainly can talk of people mechanically working through proofs in logic or mathematics, for example. However, the cognitivists seem to have lost sight of the fact that we can say this only of creatures who can also work through the proof carefully or in an inspired way, and this is not something mechanisms can do. See Peter Hacker (1990, pp. 79-80) for an expansion of this point.

[2] Of course, questions of whether a use of words is correct often do not arise. A person who always commented on one's grammar during conversation would seem pedantic. However, we can say that the utterances of competent speakers of a language do tend to conform to rules of grammar (for this is one measure of competence) but it is the explanation of this fact that invites misunderstanding, as we shall see. Those cognitivists (like Chomsky, Fodor, and, at one time, Stich) who argue for an in-built grammar, the rules of which govern the subject's linguistic production, will have to say that every thought, utterance, and, for that matter, representation is a consequence of the application of rules. The unlikely consequence is avoided by resisting the temptation 'to think that if anyone utters a sentence and *means* or *understands* it he is operating a calculus according to definite rules' (Wittgenstein, 1953, §81. I expand upon this point in section 3 below). We clearly do not operate such a calculus consciously, so the cognitivist maintains we do so unconsciously.

What I want to consider now is the question of whether it is the activity of the mechanisms, their outputs, or both, which must allow of normative evaluation. To answer this question it will be helpful to look at some forms of compliance with a rule which, for our purposes, can be subdivided into accidental compliance and non-accidental compliance. Let us begin by examining the notion of accidental compliance.

If a random number generating machine, which provides numbers for a lottery, produces the numbers '1,2,3,5,8,13' we might observe that the series complies with the formula for generating the Fibonacci series. However, since it is a random number generator the compliance with the series is wholly accidental. The machine certainly cannot be said to have applied the formula and, therefore, there is no sense in which the machine can be said to have applied it correctly, nor in which the numbers produced can be said to be correct. Here I am distinguishing between the correctness, firstly, of what the machine does and, secondly, of what the machine produces (the number string). Since the production of this number string is accidental neither it nor what is produced is correct.

Similarly, we might imagine a child reaching across a chess board for a packet of biscuits. A game is in progress and, in reaching, the child pushes a bishop diagonally forward one square. Here we can say that the movement of the piece is in compliance with the rule for moving that piece. However, the compliance is, once again, accidental. The child's action is not that of making a move in the game, so not only has she not applied a rule, but also we cannot say the movement of the piece is correct and nor, *a fortiori*, can she be correct in her acting. Since we can normatively evaluate neither the action of the child nor the output of the machine (as the notion of correctness is inapplicable), accidental compliance is insufficient for the mechanistic explanation of language use and cognition because such explanation requires that, at least, the output of cognitive mechanisms can be normatively evaluated.

Now let us turn to non-accidental compliance. Here we can distinguish between two sorts of case. The first we will look at is that of non-accidental compliance which falls short of *rule following*. If our number generating machine is designed to produce the numbers of the Fibonacci series then we can say of the output '0,1,1,2,3,5,8' both that it complies with the formula for the series and that it is correct. However, although the output, what is produced, is correct it would be misleading to say that the machine is correct, for it is its design and construction that is correct and the machine is responsible for neither.[3] To say the machine is correct can be taken to mean that it has applied the formula, but that would be a misuse of 'applied'. If this does not appear obvious one need only reflect that where there are applications of rules there can be mistakes or errors in application. If the machine produced the series '0,1,2,3,6,12' we would not say the machine had made a mistake, for making a mistake in applying a rule requires some knowledge of it and, importantly, an intention to apply it correctly. After all, one may be making a mistake when one picks up and runs with a football but not if one intends to play rugby rather than football with it. Saying the machine knows the rule and intends to use it is either to use these terms metaphorically or to talk nonsense.

Another example of non-accidental compliance will help to bring out this point. Our child has learned to play draughts and, on coming across the chess board decides to make a move on behalf of one of the absent players. She knows that this game is called 'Chess', and is different from draughts, but assumes that the difference is in the initial placing of the pieces and the objectives one has in moving them. What she holds to be constant in both games, however, is the rule for moving and taking pieces. Thus she moves the bishop as

[3] We might say that the machine is 'functioning correctly' but this can be misleading as well. It might encourage the conclusion that the machine is applying the formula correctly when it is simply doing what it is designed to do, which is; to produce number strings compliant with the formula. If it did what it was designed to do but not how it was designed to do it, it would still be 'functioning correctly' (although some of its parts might not be).

before. Here the child has made a move which complies with the rules of both chess and draughts and since she believes she has made a move in chess we can say her move is correct. It is her intention to make a move in the game that allows us to say it is a correct move. However, it would be wrong to say she has correctly applied, or followed, a rule of chess, for if she told us why she moved the bishop as she did she would do so by explaining a rule of draughts. In this case, then, it looks like the move's compliance with the rules of chess is accidental, and in a sense it is, for it is fortuitous that she decided to move the bishop and not the knight. But it would also be appropriate to say it is non-accidental because the child moved the bishop deliberately and, hence, in a manner we can judge to be correct or incorrect. Thus the child's action is correct, since the move that results is, but she has not correctly applied a rule of chess. The machine (of the last example) produced a correct number string because its designers intended it to produce numbers compliant with the formula, but it has not correctly applied the formula. The difference between the child and the machine is that the child *can* apply rules or formulæ correctly because she can intend and try to do so.

Our next, and last, example illustrates the second form of non-accidental compliance which occurs because rules are being applied, or followed, correctly. A young mathematician may be taught the Fibonacci series by being given the formula; 'After 0 and 1 proceed by adding the preceding two numbers to get the next'. When she writes '1,2,3,5,8,13' the numbers comply with the rule, they are correct, and she has applied the rule correctly. We can say this because she intends to follow the rule and can recite the formula given. If she writes '12' instead of '13' it may be because she has made a mistake in her addition, which may due to carelessness or fatigue, or she might write the wrong number through sheer recalcitrance. Whatever the reason for choosing the wrong number it is necessary, for there to be such a reason, that she intends to do something by writing it. In order to correctly or incorrectly apply a rule (and, hence, to apply a rule at all), then, it is necessary to intend, wish, desire, or try, to apply it.

2 COGNITIVE MECHANISMS AND RULE FOLLOWING

a) Which Form of Compliance?

We can now address the question of what sort of compliance must be displayed by cognitive mechanisms if their yield is to fall within the domain of normative evaluation, as it must if they are to explain language use and cognition. Should they comply with rules accidentally, then neither the output nor the activity of the mechanism can be judged correct since the accord between the rules and the outputs is a matter of pure coincidence. If this was how the output of cognitive mechanisms complied with, say, grammatical rules for a natural language then the mechanisms would have no role in a systematic explanation of linguistic competence, for the relation between adherence to grammatical rules and the competence in a language is not a matter of pure coincidence.

Note that even if a machine regularly produced a compliant output this would not imply that the compliance was not accidental. That is, should the random number generating machine begin producing *only* segments of the Fibonacci series, this would not warrant the assertion that its output was correct each time. Indeed, the lottery organisers employing it would say the opposite, for what they need are random numbers. Whether such a regularity would count as correct depends on the interests of the designers and users of the machine.

For a machine's output to comply non-accidentally with a rule and, therefore, be correct, the compliance has to be, in some way, deliberate. When a machine is designed to produce an output which accords with some rule or another, that accordance is deliberate, or intended, and it is for this reason that we can call the output correct. However, the fact that the machine is designed means that the normative evaluation of its output is dependent upon the intentions and interests of the designers and users. The rule the machine adheres to fulfils the purposes of the designers rather than the machine, and although the machine's output may be correct in the light of these purposes, its activity does not constitute a correct application of the rule because to correctly apply a rule is to be informed and, hence, guided by it. We may say that a machine is designed to follow a given rule, but this simply means that it is designed to produce outputs according with the rule rather than that it is designed to be informed by it, for the machine cannot refer to the rule prior to producing its output or make appeal to the rule in justifying its output as correct.

The foregoing point may seem irrelevant to the discussion because no one would want to say that we design the apparatus which affords our competence in a language, or our ability to think, believe, remember, and so forth. That would be absurd. However, some would argue that the apparatus *is* designed by 'Mother Nature', albeit through the process of natural selection. Dennett argues in this way (in, for example, Dennett, 1987, pp.314ff) and he is accompanied by others who favour teleological accounts of cognition, such as the evolutionary psychologists. If we accept that nature designs our cognitive apparatus to comply with norms, because the compliance enhances our chances of survival, then we would have to say that the compliance is non-accidental. But this, the first, form of non-accidental compliance, falls short of rule following, and, as I am about to argue, cognitive science needs cognitive mechanisms to be rule followers. Therefore, appeals to the intentions of Mother Nature gain no ground. The point can be supported as follows.

Even if Mother Nature designed our cognitive mechanisms to comply with rules, this would not explain what it is to apply or follow a rule. A mechanism which is designed to accord with a rule does not apply that rule because what can apply a rule can also make a *mistake* in applying it, and such a mistake presupposes an intention to apply the rule. But this is not something we can attribute to the mechanism. Furthermore, the evaluation of the mechanism's output as correct or incorrect would depend on what rule it was designed to comply with, which means that the evaluation would have to be made by looking beyond the mechanism to the intentions of Mother Nature and to her capacity to apply rules. It would be Mother Nature, rather than the cognitive mechanism, to whom reference is made in explaining the normativity of language use and cognition.

This consequence, although risible, would be methodologically satisfactory if all we required was an account of how the output of a mechanism can have a normative evaluation (an account which could be given by saying that the output conforms to the intentions of Mother Nature) but this is not all that is required. The goal of cognitive science may be expressed as that of explaining what it is to be an agent, and Mother Nature, as a designer with intentions, aims, and strategies, will be an agent. The argument I proffered at the end of the last chapter had the conclusion that, in order to attain its goal, cognitive science (particularly in the representationalist guise) would have to rely upon a naturalized account of normativity. The appeal to Mother Nature's intentions amounts to an explanation of the normativity of cognition by appeal to the cognitive capacities of an agent. But one cannot naturalize normativity by appeal to cognitive capacities unless one already has a naturalized account of these. My argument has been that the second naturalized account is dependent upon the provision of the first. Thus, the explanation of normativity presupposes the success of the explanatory enterprise for which it is required, which means there is circularity in the enterprise.

Where there is circularity in an explanatory enterprise, infinite regresses are never far away. In this case one arises as follows: The explanation of cognition is to be achieved by appeal to cognitive mechanisms and, since cognition allows of normative evaluation, the mechanisms' output must also. If the output of a cognitive mechanism was judged correct because it complied with the rules a designer (Mother Nature) intended it to comply with, then we would expect it to be the case that the designer's cognitive mechanisms are correct with regard to those same rules (for, if not it would be sheer accident that the mechanisms she designed comply with the rules). But since the only account we are offered of why a mechanism's output can be correct is that it has been designed to be so, we will need to posit a designer of the designer and, therefore, an infinite number of designers. The only way out of the regress is to say that the designer's mechanisms do not comply with rules because they were designed to, but that they do so because they *follow*, or *apply*, rules.[4]

Another way of putting the point is as follows: If cognition and linguistic competence are explicable in terms of the processing of cognitive mechanisms then these mechanisms must be rule following because only then could we evaluate normatively. For there to be an evaluation of an utterance as grammatically correct, for example, there must be something with the ability to apply grammatical rules—a word processing programme might include a 'grammar checker' but it is designed as an implementation of accepted judgements, and therefore does not make judgements, of what accords with grammatical rules. If cognitive mechanisms cannot apply such rules, and our judgements are a consequence of the workings of these mechanisms, then we could not make normative judgements, and that is belied by the facts. The appeal to natural selection as the 'designer' of our cognitive apparatus will

[4] In fact, Dennett's position is a little more sophisticated because, for him, the 'Mother Nature' idiom is an explanatory strategy at our disposal when we take the intentional stance. From this stance we can see natural selection as a process by which Mother Nature 'designs' our internal mechanisms to accord with her 'intentions'. Thus we might explain normative phenomena like reasoning in terms of the functioning of mechanisms 'designed' for this 'purpose'. From outside the intentional stance the explanation would be purely in terms of the survival of organisms, housing the mechanisms, in the environments they lived in. There is a lot that can be said about this position but I will confine myself to the following observation.

According to Dennett we can adopt the intentional, the design, or the physical stance when explaining the behaviour of organisms (Dennett, 1971). The last of these will be the one at which we reach naturalism in our explanation and it is the fact that we can do so that warrants our accepting that intentional concepts fall within the realms of scientific elucidation. However, the question that must be raised is 'are there relations of identity between phenomena individuated from different stances?' For the answer to be affirmative we would have to suppose the phenomena described in the idiom of each stance belong to the same category, for an identity relation can exist only between relata which are categorially compatible. But what category do the phenomena belong to? Do they belong to the category of intentional objects, or the category of physical objects? If the answer lights upon one of these options then it follows that they will not be individuated by the idiom appropriate to the other option (at least not prior to the identification that is in question). The consequence is that the identification of the phenomena can be made in neither idiom. The identity statement will be a 'theoretical shuttlecock' perpetually exchanged across idioms which, in the present context, means that the 'purpose' of Mother Nature *cannot* be described in the naturalist's physical idiom. We might note that the phrase 'theoretical shuttlecock' is borrowed from Ryle (1949, P.14) of whom Dennett was once a student. (Anthony Palmer gives the correct lesson from Ryle, of which my argument is an adaptation, in Palmer, forthcoming).

contribute to the explanation of the normative aspect of language use, reasoning, and action only if we can make sense of the idea that the apparatus somehow evolves—for it certainly is not programmed—not just to accord with rules, but to follow them. However, as I will argue shortly, it is misconceived to think that mechanisms can follow rules.

Before turning to the issue of rule-following mechanisms I will add another consideration to the discussion of teleological accounts of language use and cognition. If, as these accounts suppose, the rules followed in using language and in reasoning are the product of the evolutionary process, then their existence is to be explained as resulting from their fulfilment of the survival needs of our species. Thus, having rules for the use of words and, therefore, having a language, is supposedly of benefit to the species in its struggle to satisfy the exigencies forced upon it by the environment—our ability to communicate has, presumably, arisen as the result of the advantages of collaboration. A similar story may be told in giving the origins of inferential reasoning; it is, after all, helpful if one can infer both the consequences of one's actions and which actions will fulfil one's needs.

However, apart from the difficulties these stories have when it comes to specifying exactly how the transition was made from the pre-linguistic/pre-cognitive, to the linguistic/cognitive, stage—it seems to require either a point at which many people suddenly became rule followers, or an 'inventor' of rule following, and both requirements are barely intelligible—there is a distinction they ignore. It is between grammatical and inferential rules, on the one hand, and, what we may describe as, teleological rules (rules obeyed for a purpose), on the other. The distinction is best expressed by saying, as Wittgenstein was tempted to say (Wittgenstein, 1981, §320), that 'the rules of grammar are arbitrary'. That is, there is no sense in asking whether any rules of grammar we obey in speaking a language are the right ones to use, whereas there is such a question to ask of the purposive rules of, to take Wittgenstein's example, cookery. The rules a cook might obey ('always boil vegetables for at least half an hour' might be an example) can be evaluated in the light of the goals of cookery, such as to produce nutritious and tasty food, and so the question can arise as to whether adherence to a particular rule will serve such goals (which our exemplar will not). If the rules of grammar were evaluated according to whether they serve goals it would make sense to talk of mistakes, not just in their application, but also in their observance. The distinction is made clear by Guy Robinson in his paper 'Language and the Society of Others' where he observes that;

> 'If an individual or group decides to concoct names for things they have come to distinguish, they do not then have to go on to experiment to see if certain names "fit" or not. There are no mistakes lying in wait for the coiner of names, mistakes of the sort the cook or the potter may run into. (This or that sort of clay warps or cracks in firing, &c.) In linguistic matters the rule is arbitrary and the mistake is defined by the arbitrary rule and not by some naturally imposed sanction that could call the rule itself into question' (Robinson, 1992, p.337).[5]

In suggesting that rules observed in the course of language use or reasoning have evolved to fulfil our purposes, the evolutionary psychologist would have to say that such rules can be mistaken, or incorrect, in the light of these purposes. However, while we might make sense of the question as to, for example, whether the rule of double negation should always apply (for some speakers double their negatives when they clearly do not intend a positive) we cannot make sense of the question as to whether the rule is mistaken. One cannot explain the norms appropriate to language and reasoning as arising from the survival gains acquired through adherence to them. Furthermore, even if we accepted that a rule could evolve, we could not accept that a biological mechanism could follow it, as I shall now argue.

[5] Guy Robinson, in a paper dealing with the question of the scope of the private language argument, refers to *Zettel* §320 in support of the argument that a solitary individual could not be created with a linguistic *instinct*. The argument has ramifications, not only for the theories of Innatists like Chomsky and Fodor, but also for teleological semanticists like Millikan.

The differentiation between the capacity simply to do something which can be normatively evaluated and the capacity to actually apply a rule can be made by noting that the latter requires a subject able to intend, desire, or try to act in accordance with a rule. That is, rule application requires an agent. Only on this assumption can we make sense of talk of errors or mistakes, for only something that tries to follow a rule can make a mistake in doing so. We can also say that to intend, or try, to follow a rule requires that one knows the rule, and knowledge of the rule will be indicated by the ability to produce a formulation of it (though not necessarily a precise formulation), and the ability to act in compliance with the rule. These are just a few of the connections between the attribution of rule following, on the one hand, and the attribution of cognitive and conative capacities, on the other, but it is important to note that they are not connections between actions and psychological events or states, or at least that is not my claim. Rather they, what might best be described as, grammatical connections made as a result of observations on the use of the expression 'to apply a rule'. The correctness of the observations is evidenced by the ordinary way we use and justify the use of the expression and its cognates. I make this point to emphasize that fact that to say of something that it follows rules is to imply that it can do many things besides.

The foregoing suggests that applying, or following, a rule requires an agent with the capacity to display behaviour which serves as criteria for ascriptions of an intention and attempt to apply a rule and of a knowledge and understanding of it. Only when these ascriptions are warranted can something be said to have obeyed or disobeyed a rule. When we consider the behaviour which serves as criteria for such ascriptions it is clear that it is not the behaviour of a neural mechanism within an agent, or of the agent's brain, but of an agent him or herself. A neural mechanism cannot declare its intention to apply formal rules of inference to test the validity of an argument, for example, nor can it provide criteria to warrant the ascription of such an intention by picking up its textbook of logic, sketching Venn diagrams, or asking a colleague how he would translate a given sentence into standard form. We might imagine such a mechanism producing an output (when connected to, say, the vocal apparatus) which conforms to a rule, but since it is not designed to and does not intend to do so, the conformity would be purely accidental.

b) Representations of Rules

At this point the representationalist might argue as follows: The mechanism's production of an output which complies with a rule is not accidental because that rule is represented within the mechanism and, therefore, plays a causal role in the output production. The presence of the representation, *qua* neural structure, *explains* why the output conforms because it causally determines the structure of the output. It is the fact that our cognitive mechanisms contain these representations which allows us to differentiate between the correct and incorrect grammatical, or logical, forms of utterances and inscriptions we encounter. We can say that the mechanism is governed by the rule because its output is determined by it or, at least, by the representation of it. Furthermore, the representation is also causally implicated in the production of verbal behaviour, such as utterances of the form 'I intend that this action will comply with such and such a rule', as well as the behaviour indicative of an understanding of a given rule. After all, as propositional attitudes, an intention and an understanding are relations between a representation, in this case a representation of a rule, and an organism. With this in mind we can envisage the representation of the rule being operated upon within a mechanism, within an organism, to produce certain outputs, and causally constraining relations between representations (as would be the case when the rule is applied during inferences from premisses to conclusions). Thus, the representation of a rule will account for the non-accidental conformity of outputs *and* the intention to apply the rule which we have said is necessary for genuine rule following.

This argument appears to side-step the objection that mechanisms cannot *intend* to apply rules by suggesting that it is not the mechanisms which do so but the organism, or the neural system, of which the mechanism is a part. It, therefore, concedes the point that mechanisms do not intend to apply rules but still maintains that the mechanism's processing is a necessary, if not sufficient, condition of the system's rule-following capacity. However, in what follows I will try to show that the representation of a rule in a mechanism could not justify describing its output (and, eventually, the organism's output) as the result of applying that rule.

If the appeal to representations is to help in explaining how normative activity can be generated by processing within an organism, it must be because the representations are formulations of rules. Moreover, since it is to causally constrain the activity of the organism, the representation is, one must presume, a physical realization of the rule. Now, if this is so then the consequence of the appeal is a separation of the rule and its application such that the former exists, independently of any application, as a state, or structure, whilst the latter occurs as a process which ranges over that structure. Only by separating the rule from the application can sense be given to the assumption that the rule representation *determines* the application. Both aspects of the account, the reliance on the formulation of the rule and the consequent separation of the rule from any of its applications, provide grounds for rejecting it as incoherent.

To begin with, although a rule can be correctly or incorrectly formulated, an incorrect formulation of a rule is not a formulation of that rule. The apparent contradiction within this observation is removed when we consider that a person who declares an intention to formulate a given rule may make a mistake or may have an incomplete knowledge of it.

Thus, someone might formulate the rule for 'offside' in football by saying 'A player is offside only if he is nearer, than any member of the opposing team, to their goal when the ball is kicked' and a pedant might accuse him of formulating the rule incorrectly because a player is still offside if the opposing team's goalkeeper is between him and the goal. He has not, strictly, formulated the rule, but the fact that he declared his intention as that of formulating the 'offside' rule licenses our describing the sentence as an incorrect formulation. The problem created for representationalism is as follows. If the mechanism's following of a rule is a matter of its processing being constrained by a formulation of it, then the formulation must be correct (the mechanism would not be following *that* rule if it was not) but unless we have some means for individuating an incorrect formulation in the mechanism's case then there is no distinction between a correct and an incorrect formulation of the rule. But what would it be for the mechanism to have an incorrect formulation of the rule? What would be the criterion of correctness?

The answer cannot have the effect of separating the rule from the formulation, for if the rule is something other than the formulation it receives in the mechanism then it is this something, rather than the formulation, which provides the standard for evaluating the mechanism's output as correct and, hence, it is this which explains why the utterances of a subject are correct. The result is that the rule formulation in the mechanism becomes redundant in explaining the normativity of cognition. If there is no answer to the question then it seems that, whatever form the representation of the rule takes, it is correct. But then it follows that correctness is not an evaluation which can be made of a formulation of a rule in a cognitive mechanism, for what can be correct can also be incorrect. The problem is that a formulation of a rule *can* be correct, so it looks like the structures in cognitive mechanisms cannot be rule formulations.

A related problem arises from the observation that a formulation of a rule in a cognitive mechanism will have to be well-formed. If, for example, there is a mechanism which generates well-formed formulæ expressing propositions then, since 'well-formedness' is a normative evaluation applicable to formulæ, it follows that there will be norms, or rules, with respect to which that evaluation could be made. Now, since the well-formedness of the outputs of the mechanism (and, perhaps after further processing, the grammatical correctness of a subject's utterances) is to be explained by what occurs within the mechanism, it follows that the rules by which well-formed formulæ are evaluated as such will have a representation within the mechanism. The problem is clearly that *these* rules will need to be represented as well-formed formulæ which means that further representations of rules (which may be representations of the same rules) will be required to ensure that they are. An infinite regress ensues which means that there is no explanation of why the mechanism's output is well-formed because, if a representation of a rule is to govern the formation, then it can do so only after an infinite number of prior representation formations have been completed.

The problem arises on the assumption that a rule formulation must be well-formed if it is to express a rule. This assumption may not always be true depending on how liberally we apply the term 'formulation'—when applied to a gesture (which indicates that an action is correct), for example, we can see that the question of whether it is well-formed does not arise—but in the case of symbolic formulation, or more exactly, linguistic formulation, the assumption holds good. To illustrate the point let us suppose that a cognitive mechanism contains a representation of a grammatical rule expressed by the sentence 'A complete sentence includes a finite part of a verb'. We might suppose that a rule such as this would be represented in the mechanism to govern the production of grammatically well-formed formulæ, or 'messages', to be translated into the natural language. The objection just advanced relies on the affirmative answer to the question of whether this representation has to be well-formed. It would seem that the objection is avoided only by maintaining that it need not be, for otherwise the regress impends. However, the objection cannot be avoided in this way, for if the sentential representation of the rule is not well-formed then it cannot express the rule. 'A complete sentence including a finite part of a verb' is clearly not an expression of the rule cited above and, since it is not a well-formed sentence (it is incomplete because it fails to comply with the cited rule), it fails to express any rule. But, if the rule is not expressed then neither is it represented and a representation of a rule cannot be cited in explaining the production of grammatical sentences.

The difficulties also arises for representations of rules of inference. If a mechanism represents the *modus tollens* rule symbolically as $p\supset q, \sim q \vdash \sim p$, say, then the form of this representation will be of paramount importance in constraining the permissible relations between propositions filling the variable places. Thus, there is a right and a wrong way of formulating the rule, which means there is a standard of correctness here. When that standard is supposed to be provided by an internal representation, rather than the accepted practice of logicians, the formulation of the standard becomes susceptible to normative evaluation just as much as the rule formulation of which it is the standard for correctness. The result is that a lacuna is created which can only be filled by a further formulation of a rule or standard which, thereupon, creates another lacuna, and so on. The incoherence results because a formulation of a rule must be correct before it can express that rule and, since correctness is being explained as a relation between a formulation and an output, the rule formulation, as output, stands in need of another rule formulation before its correctness can be secured. Of course, the same goes for the new formulation.

The foregoing should provide adequate support for the contention that an internal representation of a rule could have no normative role to play in cognition. To be sure, a representation of a rule *can* have a normative role (it can guide action, or be used in correcting or in justifying the correctness of action, for example) but there will always be criteria for the correctness of the formulation the rule receives in the representation. By removing the representation from view—by making it internal—the criteria which apply to public rule formulations (criteria which are appealed to in the practice of applying the rule) are lost. The representationalist is then forced either to generate internal criteria, and thereby an infinite regress, or to admit that criteria for the correctness of the internal formulation would have to be public. The second option calls into question the view that internal formulations could have any normative role, because that role would include differentiating between actions that do and actions that do not accord with the rule formulated. But, since the formulation's correctness can be established only within the public domain, the differentiation between correct and incorrect actions will be similarly established and the formulation's normative role disappears. It is worth pursuing this point because it discloses both the perversity of suggesting that it is an internal cause of an action that makes it correct, and the magnitude of the mistake in trying to give rule following a causal explanation.

c) Representations as Guides to Action

Thus, we will now consider how a rule formulation, represented internally, is to guide sentence construction and inference. We can begin by recognising that the representation will not guide action as a result of its being interpreted, where the interpretation amounts to the attribution of a meaning. Firstly, talk of interpretation will fall into vicious circularity and regress, for if, by interpreting the representation, another formulation of the rule is created, then we are no closer to explaining how the rule guides action. Secondly, a rule formulation can be interpreted in any number of ways such that any action could be made to accord with the rule formulated. Consequently, no interpretation, alone, could ensure the correctness of an action performed on the strength of it. Thirdly, since cognitivists need to save appearances by making much cognitive processing unconscious, they will be distorting the use of 'interpretation' to name a process about which we have no genuine knowledge.[6]

However, the representationalist will not suppose that the rule representation guides action by virtue of its meaning. As we saw in chapter 4, the best policy will be to maintain that representations effect behaviour by virtue of causal rather than semantic properties—ultimately, to avoid content epiphenomenalism, the representationalist will need to maintain that semantic properties *are* causal. In the present context, then, the claim will be that the representation of a rule yields a correct action, or a correct 'output' (when we return to the mechanistic idiom), by causing it to be correct. That is, the output will be correct *because* it is caused by the representation of the rule, for if it is correct for some other reason the relation between its compliance with a rule and the rule representation will be accidental. In that case the representation will be irrelevant to the normative evaluation and the representationalist account of rule following will collapse.

So, the means by which an internal representation of a rule guides action must be causal. Supposing that the representation is a structure, its role in causing a mechanism's output will depend upon its becoming active in a process;[7] that is, though the structure exists over an extended period of time, it has a role in the ætiology of action only when part of a process. The cause of an output is, therefore, a process in which a structure's causal properties become active. On becoming active, however, the representation's role must be sufficient, not just for the production of an output, but also for the compliance of that output with the rule represented; for otherwise there is no sense in which it guides action

[6] All three points can be derived from Wittgenstein's *Philosophical Investigations* (1953) §§ 139-201.

[7] We can understand 'mechanism' as referring to either the entire 'cognitive' system, or a subsystem therein.

normatively.[8] The combination of the structure and the process which activates it would then be equated with the application of the rule represented.

Of course, the appeal to causation is fraught with difficulty. Not only does it fail to offer a coherent account of what it is to misapply, or to make a mistake in applying, a rule, but it also fails to deliver what it promises, *viz.* a mechanistic account of what it is to apply a rule. Let us uncover these failings.

We have just established that, in order to secure a non-accidental relation between the rule representation and a compliant output, the representationalist needs to cast that relation as causal, rather than semantic. In addition, he or she will have to suppose that the representation will determine the output as correct by virtue of the fact that, when activated, its structure will be sufficient for a correct output, as effect. So, if we understand the activation of the structure as a process then we can say that when the representation, R, figures in the process, P, the correct output, O, will result of necessity. Now, the question we must ask is, if R+P is to amount to an application of the rule then what will count as a misapplication of the rule? To misapply a rule is to use it wrongly or inappropriately, or to make a mistake in applying it. In any case, the rule has a role in identifying an action as a result of a misapplication, for the action is identified as the result of a misapplication of *that rule.* So, R, the representation of the rule, will play a role in the misapplication. However, that means that R is not jointly sufficient for a correct output after all, for R must help cause an incorrect output when it is part of a process of misapplication.

The obvious way round this is to suggest that when the output is incorrect, the process that activates R is not P. P may be the process in which R is activated to cause correct

[8] One might think of the representation in the mechanism as a key in the lock of a door. When the key is turned it is the structure of the key, its shape, that necessitates the unlocking of the door. Similarly, when the representation is activated, it is its structure that necessitates the compliance of the output. This implies that had the process of turning, or of activating, occurred with a different key, or representation, it might have been insufficient for the unlocking or the compliance of the output. It should be added that in establishing these types of causes as sufficient for their effects, a *ceteris paribus* clause would need to be introduced to the statement generalizing the relation. Thus, the turning of a certain key will necessitate the

outputs but, if P is replaced by another process, P*, then R may produce an incorrect output. Thus, when R is activated by P* we have an instance of misapplication. This would be all well and good if it was not for the fact that by allowing for a range of processes in which R becomes efficacious we allow for a range of outputs not all of which accord with the rule. The problem is that we now require a criterion for their correctness other than their having been jointly caused by R. We cannot obtain the criterion by examining the processes that produce the outputs for they will all be processes activating R. So, we cannot tell which process is an application, or a misapplication of the rule, firstly, unless we already know which output is correct, which we do not, or, secondly, unless there is some criterion of correctness of outputs independent of R's causal relations. But if there is some independent criterion of an output's correctness, such as that it leads to action that concords with the public practice of applying the rule, then it is this that will explain the correctness of the action and not an internal representation of the rule. In other words, R will not, causally, 'guide' the application of the rule because R cannot determine whether or not what it causes complies with the rule.

Put another way, if the criterion, appeal to which establishes the correctness or incorrectness of the output, is independent of any process by which R necessitates its effect then the relation of R+P to the correctness of O is purely accidental. Thus, it might be that another internal structure, S, when activated by P, caused O,[9] yet there would be no reason to deny O's correctness on account of the fact that it was not necessitated by R, for the criterion of O's correctness is independent of R, and the process by which it determines an effect. But if the relation between R, and P, and the *correctness* of O is accidental then R+P is not an application of the rule, and, *a fortiori*, R does not guide the application of the rule.

unlocking of the door, *in the absence of factors inhibiting the workings of the locking mechanism.*

[9] This could happen because, as a cause, R+P is sufficient but not necessary for O.

Note that the claim that R and P are *nomologically* related to O does not help here. It might be claimed that a relation of R and P to O is an instance of a nomic regularity and that R+P is, therefore, *necessary* as well as sufficient for O.[10] This means that the relation of R+P to the correctness of the output would not be accidental because O could not have been caused by another process, involving another structure. This will not help because R could still be activated by other processes, by P* for example, in which case R+P* will produce another, incorrect, output, O*. We would, therefore, have two nomic relations covered by the law L, for (R+P) ≡ O, and the law L*, for (R+P*) ≡ O*. Now, since R is present in both relations it cannot be the determinant of the correctness of O, so the nomic statement, the formulation of the law, will contain nothing which differentiates between the causal antecedents such that one is necessary and sufficient for the output's being *correct*. The difference between P and P* is not one that can differentiate between correct and incorrect outputs because P could be distinguished as the process leading to the correct output only after it has been established which output *is* correct. This means that the only way to ascertain which of L and L* describes a nomologically necessary relation between a rule representation and an output which complies with the rule, will be to light upon a criterion, independent of the laws, for the correctness of O. But if the criterion is independent of the laws then, again, the fact that R+P always produced an output in accordance with the rule would be a matter of happenstance, for it could have been R+P* or S+P that did so.

The best manoeuvre might be to jettison the claim that a *structure* represents a rule. Thus, by claiming that rules are represented by *processes* rather than structures, the representationalist might appear to bypass many of the trouble spots created by allowing a separation of structure and process. In doing so, however, he will have taken a road that

[10] Here I am following Davidson who maintains that causal laws have the form of a conjunction of statements giving causes as necessary and sufficient conditions for their effects (Davidson, 1967, p.158).

speeds him away from any account of rule following. In the mechanistic picture the rule representation, as structure, was supposed to guide the process of application so that following a rule will amount to a process in which a structure causally determines a correct output. The avoidance of incoherence in this supposition, by taking the representation to be a process itself, would be achieved only by identifying the representation with the application.

However, although there is reason to say a rule is indistinguishable from its applications—there is nothing to a rule beyond a rule-following practice—a *formulation* of a rule *can* be distinguished from its application, for it can be used to show that an action is a misapplication. Indeed, if the process representing the rule was the process of *application* then, on this account, there would be no process, representing the rule, that could be identified as a *misapplication* of the rule. The claim that a process represents a rule, therefore, will be viable only if that process can be distinguished from a further process of application, thereby reintroducing the problems faced by the structural account (for whether the first process determines the correctness of the mechanism's output will depend upon whether the second process preserves the 'normative' determination of the first; but which of the possible second processes does so would have to be decided by recourse to the rule, not its representation). Besides, while the formulating of a rule might be a process (as when one writes down a rule of grammar) the formulation itself is not one.[11]

I dare say that the representationalist account of rule following could undergo further contortions in the struggle to forge a nomic link between a representation of a rule and an action which complies with it, but whatever shape representationalism takes it will be unable to offer a consistent account of rule following as a mechanistic process. That mechanistic explanation is incommensurable with normative explanation is evidenced by

[11] I hope it is clear that I have been using 'formulation' to designate the result of formulating rather than the act or process of formulation.

the fact that when the latter is distorted to fit the former it cannot retain integrity. For example, the rule representation is supposed to govern or guide the mechanism's processing by causally determining the correctness of its output. But, if a representation of a rule determined the correctness of a mechanism's output, then there would be no possibility of *misapplying* a rule. Thus the mechanistic explanation, in failing to allow for this possibility, cannot measure up to the normative explanation. In fact, the mechanistic explanation cannot even secure the antecedent condition that a rule representation determines the correctness of an output.

The representationalist is right to separate the representation of a rule from its application but, in locating the representation within an organism, in burdening the representation with the normative role of the rule, and in construing this role as that of determining an action as correct, he requires a 'super-mechanism' constructed from an immutable substance such that its processing can never deviate. No mechanism can fulfil this description, whether it be neural or otherwise, which means that there can be variance in the processes in which a rule representing structure would become causally active. If *they* can vary then so can the outputs of the mechanism, and it is facts such as this that force asunder normative and causal explanation. This is because the consequence of the under-determination, of the correctness of an output by a representation, is that the arbitration between correct and incorrect outputs must be separated from the causal role of the representation.[12]

[12] In thinking of a representation of a rule as determining in advance, by virtue of its nomological connections, all possible applications of the rule the representationalist is succumbing to the temptation Wittgenstein characterizes by his 'rules as rails' analogy (Wittgenstein, 1953, §§218-221 and MS 165, 83). In viewing a rule (representation) as determining the possible actions of the rule follower in the way that a track determines the possible positions of a locomotive, the theorist confuses a normative connection (between a rule and an action) with a causal one. That this is a mistake becomes apparent when we remember that a locomotive can become derailed. There is no analogous situation for the rule follower, for, if rules were analogous to rails, a deviation from the course laid down by the rule would be either a misapplication, in which case the rule failed to determine the application, or an application, in which case the rule is followed even when one's action fails to comply with it. Both the first horn of the dilemma (the failure of the rule representation to determine as correct the action it causes) and the second (the identification of an incorrect action as the result of following the rule) *should be* unacceptable to the representationalist. The conclusion is that a representation of a rule cannot causally determine an application.

The relevance of Wittgenstein's remarks to causal accounts of cognition is indicated by Baker and Hacker (1985, pp.212&213) and Thornton (1998, p.45). A line of argument offered by Baker and Hacker in *Language, Sense and Nonsense* provided the stimulus for a good deal of what I say in this chapter(see Baker and Hacker, 1984, pp.294-299).

By making the relation of the representation to the output nomic the possibility of correct outputs being caused by other structures is removed (since the representation and a particular process type are necessary and sufficient for the correct output), but since the representation is separate from the process of application the possibility of its occurrence in other processes is not removed, and the problem of arbitration reappears. The upshot is that if the causal role of the representation is to correspond to the following of a rule, then, since that role is variable, in *following the rule* the mechanism can produce outputs which *fail to accord with it*. But to have followed, or applied (rather than misapplied), a rule one *must* have acted in accordance with it. The mechanistic story undertakes the hopeless task of rendering that 'must' as a mark of causal determination when its force is, in fact, normative. Let us, briefly, consider this point.

There is a sense in which we can talk about a rule determining what counts as an application. The determination here is captured by the condition that one *must* act in accordance with the rule *if* one is to be described as having applied it. However, the determination is not of the application by the rule (for a *rule* is not a condition of its application) but of the correct use of the expression 'applies a rule' by another, 'accords with the rule'. Similarly, one may be constrained by rules insofar as, for example, *if* one wishes to play chess then one *must* obey its rules (for if one strays from them one is no longer playing chess). Here, again, the rules of chess (or their formulations) are not acting upon one's performance—no such action could account for the 'must'—but a relation is being stated between the applicability of two forms of description, viz., '*x* is obeying the rules of chess', and '*x* is playing chess'.[13] To suppose that the satisfaction conditions of one expression *must* bring about those of the other is to seek an explanation of the 'must', beyond the use of the expressions, in the relations between things, and thus to treat it as

[13] Cheating, of course, is 'not playing the game'

indicating nomological necessity. This is wrong because the force of the 'must' is normative, for we are using it to distinguish between the correct and incorrect usage of expressions. My hope is that the arguments proffered show that the normative cannot be reduced to, or explained by, the nomological.

CHAPTER 7

EXPLAINING COGNITION

The fact that normativity cannot be accounted for in the naturalistic idiom has far reaching consequences. Indeed, this fact alone suffices to show that there cannot be a science of cognition. The following summary of the enterprise of cognitive science brings this point into full relief:

Philosophers of cognitive science presuppose that among our psychological vocabulary there are terms which refer (or, in the Eliminativist's view, purportedly refer) to cognitive states and processes. Not the least because viable scientific entities need to be quantifiable, these states and processes are taken to be states and processes either of the brain, or realized in the brain (for, as we saw, Functionalists cast beliefs, for example, as functional states which could be realized in any suitable medium). Naturally as physical states and processes they interact with one another and the organism housing them, in virtue of their causal properties thus satisfying another desideratum of a natural science, *viz.* that relations between entities individuated in the taxonomy of the science be subsumed under causal laws.

However, those who consider themselves Realists also see cognitive states and processes as semantically laden and believe that the relations they bear, to each other and to an organism's behaviour and environment, are semantic. In other words, they are relations involving the contents of the states and processes such that the relations will be constrained by those contents; by what the states and processes are about. This constraint can be loosely described as rational in that its exercise results in such phenomena as practical reasoning (which, even when erroneous, requires that beliefs, desires, thoughts, and the like, be semantically related in some way) and rational behaviour. Of course, we should note that 'rationality' is a normative concept. That is, provided we are dealing with rational, as opposed to non-rational, creatures we are able to evaluate what they do and say as rational or irrational, and we are able to justify these evaluations by appeal to norms, standards, and rules of reasoning.

The Realist's picture culminates in the claim that *causally* related internal states and processes have contents which bring them into *rational* (and, therefore, normative) relations with each other, with behaviour, and with the environment. Thus, by depicting rationality as a form of causality, cognitive science appears to bring cognition and content within the scope of naturalistic explanation. The possibility of identifying beliefs, thoughts, and the like, with neural states and processes is taken for granted among the philosophers we have been considering, but we should note that they are helping themselves to another form of identification; for they require an identification of *normative* with *causal* relations. While I believe that the first form of identification is entirely misconceived, it is the second with which I have been concerned. Clearly, the two are related as not only is the first presupposed by the second, but without the second the first loses all intelligibility. In addition, those who reach their physicalism through Functionalist premisses will require the second identification if they are to argue that mental states are physical *because* their relations to each other and behaviour—relations individuated in virtue of their *rationality*—are causal relations individuating a neural state.

In the next two sections I will try to draw out some of the ramifications, for Realism, of a rejection of this identification of normative with causal relations; a rejection I believe was justified by the previous chapter. In the third section I will reconsider Connectionism in the light of the arguments of the last three chapters and push for the conclusion that whatever Connectionist modelling might explain, it will not be cognitive activity and language use.

Eliminativism has been largely absent from the discussion in the last three chapters but in the eighth, and last, chapter I will try to show why its central premiss, that the psychological vocabulary is theoretical, and its conclusion, that the vocabulary should be replaced, are wrong. Let us proceed, then, by returning to the issue of mental, representational, content which was the focal point of chapter 5. Beginning at the end of the chapter we can formulate a conclusion as to the possibility of a naturalistic semantics for psychology based on the notion of a conceptual, or inferential, role.

1 MEANING AND NORMATIVITY

The central claims of Conceptual Role Semantics are that the meaning of an expression is a matter of its conceptual role, and that conceptual role is a matter of causal role. The point I wished to highlight was that the conceptual role of a symbol could be appealed to in justifying the assignment to it of a meaning only on the assumption that there is a normative framework in which it has its role. A symbol's relations to other symbols in sentential and inferential contexts would afford it a determinate meaning only if there was a justification for saying it had one, rather than another, meaning. The justification for claiming the meaning we assign is the correct one is that only if the symbol had that meaning would the inference in which it occurs be valid. That is, only if the symbol means '...' will the role it plays bring the inference into conformity with rules of inference, or, put another way, only that meaning would render the use of the symbol correct.

Since the conceptual role of a symbol is a matter of its causal role, and since, if the symbol is to have meaning, the role of the symbol needs to be in accordance with rules of use, it follows that the causal role of the symbol accords with rules. Note that the accordance of role with rules cannot be accidental because, after all, the relation between a symbol's role in an inferential context and its meaning is not accidental - if 'and' did not serve as, or mean, a conjunctive then the inference from 'John is tired and emotional' to 'John is tired' would not accord with norms of inference. Thus, the accordance of role and rules will have to be non-accidental. Furthermore, the role has to accord with the rules because the rules are being followed; for if the accordance was non-accidental because it was designed to be so, the account of the meaning of the symbol would have to import the notion of a designer, which is to abdicate naturalism. The requirement, therefore, is that the causal relations into which a symbol enters constitute applications of rules. The medium in which these causal relations occur, the brain, will need to be a follower of rules for the use of symbols in inferential contexts.

The arguments presented in the last chapter should show that the brain, or a part of the brain, cannot be a follower of rules of inference, or of any other kind. If it could it would make sense to talk of the brain making mistakes or errors in its application of rules and, as I have stressed, this only makes sense if it also makes sense to talk of the brain intending, or trying, to apply a rule. Such talk cannot be made coherent by accounting for the intention to follow a rule in terms of an internal formulation which causally constrains the brain's outputs such that they conform to the rule. Not only will there be no account of what it is for such a formulation to be correct, but the dilemma will emerge by which either the brain will be incapable of misapplying a rule, or will apply the rule even when its output fails to comply with it.[1] The conclusion, that the brain does not apply rules, when directed at Conceptual Role Semantics, indicates that the causal role of a symbol could bear no relation to the meaning that symbol has in an inferential context. Conceptual Role Semantics,

[1] Block assures us that he is not committed to 'rules for reasoning being themselves represented' and observes that 'in computers we have examples of symbol manipulators many of whose symbol-manipulating "rules" are *implicit* in the way the hardware works' (Block, 1986, p.107, my italics). It is difficult to understand this use of 'implicit'. We might, for example, take Dennett's line and say that something is implicit if it is logically entailed by what is explicit (see Dennett, 1983, p.216). However, even if we ignore the fact that prior to interpretation by a subject it makes no sense to speak of the workings of a computing device as logically entailing anything, a rule which may be entailed by a set of actions is not necessarily a rule that is being *followed*, and my contention is that a naturalistic explanation of the normativity of reasoning requires a naturalistic explanation of rule *following*.

therefore, cannot provide a coherent account of the content of items whether internal or external.

In arguing that Conceptual Role Semantics presupposes a normative framework which cannot be naturalized I am not suggesting that such a framework *is* the source of linguistic meaning. Rather I am indicating that the determination of meaning by reference to conceptual role *would require* such a framework, for the determination of what a symbol meant in a given inferential context would have to involve appeal to the norms and rules applicable to that context. That a symbol means 'and', for example, would be determined by the fact that that is what it *should* mean if the inference concerned is valid. The reason why I do not claim that the normative framework is the source of a word's meaning is that this claim brings with it a conception of language as a calculus. That is, it encourages the conception of a system of predetermined rules according to which words are combined licitly, to produce sense, or illicitly, to produce nonsense. By uncovering what is wrong with this conception we gain further perspectives on the faulty reasoning responsible for many attempts to explain linguistic competence.

Linguistic rules, if they are to fix the role of words, must do so by determining in advance those combinations which are licit and those which are not. That is, the rules must lay down in advance all permissible locutions in the way the structure of an idealized machine—a machine invulnerable to mutation—will determine all possible results of its machinations.[2] There is a good deal to be said against this consequence of the conception of language as a calculus.

To begin with, it is not difficult to envisage situations in which we would not know what it does and what it does not make sense to say, not because we have imperfect knowledge of

[2] This comparison between a rule and an idealized machine is drawn by Wittgenstein (1953, §§193&194)

the rules for the use of the words we want to use, but because no rules can anticipate every possible situation in which we want to use those words. At section 80 of the *Philosophical Investigations* Wittgenstein imagines the case of a chair which, time and again, disappears and reappears. The question he raises is whether we are to use the word 'chair' here—for we may wonder if 'illusion' is the correct word to use—or, more exactly, whether we have a rule for the use of 'chair' which covers such situations. It is difficult to see how there could be such a rule, but its absence would not render the word meaningless in ordinary situations and it is this fact that points to the rejection of the thought that unless a word's use is completely determined by rules it cannot have a determinate meaning. Furthermore, the thought that there *must* be rules completely determining a word's use is spurious because no set of rules could serve this function. There can always be doubt as to how, and whether, to apply any of those rules. Wittgenstein makes the point as follows;

> 'I said that the application of a word is not everywhere bounded by rules.[3] But what does a game look like that is everywhere bounded by rules? whose rules never let a doubt creep in, but stop up all the cracks where it might?—Can't we imagine a rule determining the application of a rule, and a doubt which *it* removes—and so on?' (ibid., §84)

The 'and so on' indicates the infinite regress lurking behind the assumption that a word's legitimate use, or its meaning, can be completely determined by a set of rules. We can always doubt that any particular use is a legitimate one, that on any occasion we are doing what the rules for the use of that word demand, but the doubt cannot be removed by the introduction of further rules for the correct application of the original ones, for the new rules can be accompanied by new doubts.

The conclusion should be that what makes the use of a word legitimate, or what makes it meaningful, is not the fact that the use is constrained by rules precisely determining its role.

[3] See, for example, §68.

The consequence, however, is not that we can never give a determinate meaning to a word in any of its uses. The ability to imagine a doubt about any one of those uses does not entail that we do so doubt, as Wittgenstein points out later in the section just quoted. Rather, the consequence is that what makes the meaning of a word determinate is largely a matter of the situation in which it is used and its fulfilment of the purpose for which it is used.[4] Against the objection that such surroundings of a word's use do not give the exactness implicit in the notion of determinacy I offer another passage from the *Investigations*;

> 'If I tell someone "Stand roughly here"—may not this explanation work perfectly? And cannot every other one fail too?
>
> But isn't it an inexact explanation?—Yes; why shouldn't we call it "inexact"? Only let us understand what "inexact" means. For it does not mean "unusable". And let us consider what we call an "exact" explanation in contrast with this one. Perhaps something like drawing a chalk line round an area? Here it strikes us at once that the line has breadth. So a colour edge would be more exact. But has this exactness still got a function here: isn't the engine idling?' (ibid., §88)

The directive 'Stand roughly here' represents the degree of exactitude with which we might explain the meaning of many of our expressions. The requirement that for a word to have a determinate meaning there must be rules constraining its use by, as it were, marking an exact position for its meaningful insertion into any context, is a requirement that is both never to be met and idle. As long as the role, for which the word is intended, is fulfilled there is no need for further constraints upon its usage.

Adherents of Conceptual Role Semantics, in maintaining that natural language sentences inherit their semantic properties from their internal counterparts, require that words of the internal language have their meaning in virtue of a role predetermined by rules. The option is not open for them to allow that, in a given circumstance, a word can be spoken

[4] Note that our using words for a purpose does not entail that there is a purpose behind the rules for the use of those words. Analogously, pursuing a particular strategy in a game of chess does not lend strategic value to the rules of chess. This point relates to what I said about the difference between linguistic and biological norms in the previous chapter (section 2 a)).

meaningfully even though there is no rule determining its use in that circumstance. In such a circumstance the word may be said to play a role, but they cannot allow that what that role is will be shown by the surroundings (such as the manifest intentions of the speaker, the social customs in which verbal exchanges are embedded, the form of the relationship between speaker and listener, and more obviously, the *relevant* aspect of the environment in which the exchange takes place) for these are the surroundings of the spoken word, not the Mentalese equivalent from which it is supposed to inherit its meaning. That is, the role in question is the role of the spoken word rather than the internal equivalent. So in lieu of any surroundings which might show what meaning is being given to a word, the Conceptual Role Semanticist must appeal to interrelations of items of the internal lexicon, and the rules binding these, as the determinants of their role and, therefore, meaning. That is, he must appeal to a calculus of rules for the use of Mentalese words.

Of course, such a calculus could not determine the meaning of a word in a given context, for doubts can be raised both as to whether that context is covered by the rules, and as to how the rules are to be applied in any context. Furthermore, the need for such a calculus does not arise in ordinary contexts. Here words are understood not because we apply an exact framework of rules defining their meanings, but because they are thought of and acted upon in ways appropriate to the uses they are given.

Thus, what are taken, within cognitive science, to be causal consequences of linguistic understanding are, in fact, the criteria for attributing that understanding. A similar point can be made regarding the application of rules. What the cognitivist takes to be the causal consequences, and therefore separable from, an application of a rule are, in fact, the criteria for saying a rule has been applied. It is the failure to grasp this point that encourages the assumption that applying a rule is a process in which the rule constrains the consequences of its application. As we saw in the last chapter, thinking of rules as constraining, by causally determining, their applications leaves no room for mistakes in applying them. The putative internal process of applying a rule, in other words, cannot be the criterion of the rule's having been applied, so the process has no part to play in explaining what it is to apply a rule.

These considerations speak against the appeal to the framework of predetermined rules, constraining the role of words and, thereby, defining their meanings, which I have suggested is presupposed by the Conceptual Role Semanticist. They also speak against the picture of communication as the transference of 'messages' from one representational system to another. This is the picture in which a sentence is formed by combining tokens of internal symbol types, according to an internal grammar (a system of predetermined rules) whereupon the combinatorial structure is translated into a perceptible object (a 'wave form'). Reversing the process—replacing combination with analysis—is supposed to explain how the sentence is understood. This is, of course, the picture of communication that we looked at in the first chapter when considering Fodor's Language of Thought Hypothesis.

2 THE LANGUAGE OF THOUGHT

If what I said in chapter 5 is correct we have firm grounds for rejecting the claims that there are internal meaning atoms and, therefore, that representational, or judgeable, contents can accrue from these. We might add that if the analysis of sentences into constituents yields meaning*less* atoms then, even if such a process occurred internally, it would not shed any light upon what it is to understand the meaning of a sentence.[5] In the last chapter I offered further grounds for rejecting the claims by raising objections to the assumption that there is

[5] There is a related line of objection that one might pursue here. It concerns the problem of analysis which has troubled philosophers of logic from Bradley and Frege onwards. The problem is, in short, that in analysing a proposition into constituent parts we destroy the unity which was essential to its expressing a sense, for not only are we left with a list of components which

an internal system of rules, or generative grammar, for the combination of meaning atoms to form the representations supposed to express the contents of the propositional attitudes and to cause behaviour appropriate to them. In positing mechanisms, 'compilers', for the rule governed generation and decomposition of internal 'messages' Fodor implicitly assumes that we can make sense of rule following mechanisms. If, as I have argued, mechanisms, although they might produce outputs that accord with rules, cannot *follow* rules, then Fodor's picture of communication, and his corresponding account of linguistic competence, cannot be accepted. In addition, we cannot explain cognitive concepts like 'reasoning', 'calculating', and 'inferring', as concepts of processes by which internal representations of rules are applied to internal sentences; for, if the explanation is to be naturalistic, the relations between rules and sentences would have to be causal and causal relations cannot be appealed to in explaining the normativity attaching to these concepts.

In fact, upon acceptance of the incompatibility of causal and normative explanation, the whole body of Fodor's theoretical claims disintegrates. His claim that we *learn* a natural language by making and confirming hypotheses, within the medium of an innate language of thought, about the truth conditions of natural language predicates, requires that hypothesizing is a computational process ranging over states of an organism and, therefore, that the confirmation of a hypothesis be explicable in terms of a causal process. But causal processes cannot confirm anything because *they* cannot provide a criterion by which to judge whether causal outcomes are correct. The complementary explanation of

no longer expresses anything that might be judged true or false, but also we will not find anything in that list to account for the unity of the proposition prior to analysis. Identifying, within that list, a relational item which purportedly brings the other items together will not help because the list will have to be supplemented by a further relational item which relates the first to its relata.

Frege's solution was to suggest that amongst the items of the list we find concepts, incomplete functions which are completed only when a term fills their argument place, at which point they acquire a truth value and, hence, the status of a proposition. Unfortunately, the nature of concepts, or functions, would seem to be such that they cannot be the subject of propositions and cannot, therefore, be individuated as a residue of analysis. Fodor's explanation of *understanding* a proposition, expressed by a sentence, as a process of internal analysis inherits the, in my view insurmountable, difficulties attaching to this conception of logical analysis. See Palmer, 1988, for insights into the problem of analysis.

understanding a natural language as having internalized a theory of meaning (a truth definition for that language's predicates) becomes untenable when we remember that the employment of that theory amounts to the application of truth rules by compiling mechanisms. To understand a predicate of a natural language the relevant compiling mechanism must represent the truth rule for that predicate of the form: [*Py*] is true (in *L*) iff *Gx*; where *L* is the object (natural) language and the rule is expressed in the metalanguage, the language of thought. Thus, the application of that rule is effected mechanically. But then the explanation of what it is to understand predicates of a natural language falls foul of the objection that mechanisms do not apply rules.

Fodor argues that understanding the language of thought differs from understanding a natural language in that the former does not require a truth definition because;

> 'the machine language differs from the input/output language in that its formulæ correspond directly to computationally relevant physical states and operations of the machine: The *physics of the machine* thus guarantees that the sequence of states and operations it runs through in the course of its computations respect the semantic constraints on formulæ in its internal language. What takes the place of a truth definition for the machine language is simply the engineering principles which guarantee this correspondence' (Fodor, 1975, p.66, my italics).

If by 'semantic constraints' Fodor means the need to use and combine expressions correctly—in accordance with norms or rules—then the physics of the machine cannot guarantee that rules are respected because the way in which a physical state or process can determine a machine's output is not one in which a rule determines what accords with it. Indeed, in talking of rules as *determining* that which accords with them one teeters on the edge of a philosophical precipice. As a rule for the application of a natural (input/output) language predicate, a truth rule cannot be applied by a machine, but neither, in the case of the machine language, can the correspondence between uses of predicates and the *correct* uses of predicates be *guaranteed* by the machine's 'engineering principles'.

The passage quoted above occurs as part of Fodor's reply to the infinite regress argument. As we saw, in chapter 4, his strategy is to claim that the role of internal representations in the ætiology of behaviour is afforded by their causal, rather than semantic, properties so that it is not necessary for a representation to be interpreted or understood in order for it to be a contributory cause of behaviour. However, in the light of the recent arguments, the strategy is no longer open to him. For Fodor, beliefs, for example, are computational relations between representations and the organisms housing them. Verbal behaviour which results from the having of a belief that it is raining, then, will be a product of such a relation and, given that representations are efficacious by virtue of their (physical) causal properties, it follows that verbal behaviour will be a product of a causal relation.

If we accept this picture, for the moment, then we must also accept that there will be a semantic relation between the cause (the representation) and the effect (the behaviour). The belief's representational content 'It is raining' will have a semantic relation to the behaviour expressing the belief, such as the utterance 'It is raining'. Put simply the representation and the utterance have the same content; they have the same meaning. Furthermore, given that the utterance is an expression of the belief, the relation is one susceptible to normative evaluation; the expression of the belief by the utterance 'It is raining' is correct while its expression by 'Someone is standing on my foot' is not (cf. ibid., p.73 note 13). For Fodor it is a theoretical constraint on any computational model of cognition that its derivation of the content of representational states yields a matching, between belief states and their causal consequences, which corresponds to the relations between the content of those states and their logical consequences (which, we might assume, will include the behaviour they cause, or the content of the description of that behaviour) (see Fodor, ibid., p.198-200 and chapter 1 above).

So, the picture requires that the appropriateness of behaviour to a belief, and *vice versa*, is guaranteed by the causal relations between representations and behaviour. However, since that appropriateness is a matter of whether the behaviour can be taken as a correct expression of the belief the requirement is that the causal relations guarantee the correctness of an organisms output. But causal relations cannot do this, for if the behavioural effects *were* necessitated by the representational causes then we would be forced to say that the effects are correct regardless of what meaning we attribute to them. The consequence is that an evaluation of correctness is inapplicable here so both logical and illogical consequences of belief states are ruled out.[6] But then one could express one's belief that it is raining by saying 'Someone is standing on my foot'. Fodor's suggestion that, should this be a consequence of our derivation of the content of a representational state, we would revise our computational model does not offer a way out because such a consequence can occur on any model. The problem is that so long as we envisage an extensional relation between belief and behaviour and explain the relation causally, the normativity accompanying that relation will be inexplicable. But without the normativity we would have to accept that saying someone is standing on one's foot *is* expressing a belief about the weather.

In chapter 4 I argued that Fodor cannot avoid the infinite regress argument if he wants to maintain that brain states can be interpreted as representational, which he must in order to support the thesis that there is a language of thought. When the cognitive scientist interprets a brain state as a representation this yields an interpretation of that state which, for the propositional attitude Realist, amounts to the creation of another representation. It might seem that a way out of the regress is to say that this representation does not stand in need of interpretation in order to acquire its representational status. The behaviour of the cognitive

[6] On the other hand, if the representation did not necessitate the behaviour and the causal relation between them did not guarantee the correctness of the outcome, then the computational model will fail to account for such truisms as that one's belief that it is raining is the *reason* why one says it is. The rationality of action and behaviour, in other words will be not be

scientist, for example, will indicate that he has interpreted the brain state as meaning *P* and we can assume that his internal state has the representational content 'the brain state means "*P*"'. However, if the causal relations between representational states and behaviour will not support the normative judgements required—so that we cannot give the content of the representational state on the grounds that that content would be *correct*, or *appropriate*, given his behaviour—then the scientist's behaviour does not give grounds for inferring that he is in a representational state (although his *bodily movement* might justify an inference as to the *state of his brain*).

After attempting to disarm the infinite regress argument in *The Language of Thought*, Fodor turns his attention to Wittgenstein's argument against a private language (see, in particular, Wittgenstein, 1953, §258). As he sees it, 'Wittgenstein is basically concerned to show that no definite sense attaches to the notion of a term in a private language being used coherently (as opposed, eg., to being used at random)' (Fodor, 1975, p.69). He takes the argument to be that a private language lacks coherence because it lacks public criteria for the correct application of those terms. The incoherence follows because there will be no evidence to show that a term, purported to name a private object or event, is being applied correctly in any given instance. If there is evidence then it should be publicly accessible, and if there is nothing serving as evidence then there will be no difference between using the term coherently and using it at random, 'and a term that may be used at random is no term at all. And a language without terms is no language at all' (ibid.).

As Fodor sees it the argument poses a problem for the Language of Thought hypothesis in so far as its terms, although they *do not* refer to things accessible only to the speaker, *do* lack public criteria for their applicability (ibid., pp.69&70). The generation, from internal symbols, of a formula in the language of thought is not governed by public conventions for

among the explananda of computational psychology.

the use of those symbols but by causal constraints built into the compiling mechanism. It seems that Fodor sees his task as that of showing how the coherence in the use of expressions, which in a natural language is a matter of accordance between use and conventions for use, can be explained, for the language of thought, in terms of nomological necessity (ibid., pp.71-79). I shall sketch an outline of Fodor's proposal for accomplishing the task.

In the manner of many a misinterpreter of Wittgenstein, Fodor suggests that he saw the coherent use of public language terms as 'controlled by the ...*conventions* [which] relate the terms (in many different ways) to paradigm public situations'(ibid., p.71, my italics). This is supposed to imply that Wittgenstein believed that 'a term is coherently employed when its use is controlled (in the right sort of ways) by facts about the world' and, as Fodor sees it, the belief is wrong because what matters for coherence is not that the world accords with a speaker's use of a term but that his or her beliefs about the world do so. On this assumption he gives the condition for coherent representation in a public language in the following unwieldy formulation (ibid., p.78):

(C) (S uses 'a is F' to represent *a*'s being *F*) just in case ((S believes that *a* is *F* just in case S assents to 'a is F') is conventional)

So, for example, a speaker, S, means that some object is red by 'This object is red', if and only if it is a convention that he believes that it is red, if and only if he says 'This object is red', or agrees when that sentence is uttered.

The idea seems to be that knowing that S believes that *a* is *F* gives us grounds for saying that he uses his terms coherently when he says 'a is F'. This is because we will know that S uses these terms to state what he believes is a fact, and we know that the convention is to use those words to state that fact. Thus, we will know that when S says 'a is F' he means that *a* is *F* and, as Fodor sees it, that is enough for coherence in language use.

We might note that the contention that the use of words is governed by convention is problematic because it implies explicit agreement in use. The fact that language speakers seldom seek, or need, to reach such agreement is attested by the inability of most speakers to formulate governing conventions for most of the words they use. Furthermore, such explicit formulations could not determine a word's use unless they emerged from an established practice of using that word for, as we noticed in section 1, a system of pre-established rule formulations could not determine a word's meaning (or its coherent use). Of course, an established practice will be a manifestation of a form of agreement, only it is not an agreement based on the convening of speakers for the purposes of formulating rules. I will say more about this shortly.

To forestall confusion we should recall that (C) is not intended to outlaw a private language, the terms of which refer to private objects or sense data. With regard to a, putative, private object such as a sensation, a belief that that sensation is what one normally calls 'S', even if that was necessary and sufficient for ones declaration 'This is "S"', would not guarantee that one is using 'S' to name the same sensation one had undertaken to give that name, for believing one is doing so is not the same as doing so. In other words, in the case of the necessarily private language, the relation between the sign and what it refers to is the important one because the corrigibility of declarative sentences, taken for granted in public language, is noticeably missing in the case of the private one.

Thus, condition (C) does not appear to have any relevance to Wittgenstein's private language argument. However, we should acknowledge that Fodor sees that the language of thought would not be a necessarily private language, for its terms need not name private objects, and the system of representation, like the computer's machine code, is, in principle, publicly accessible (ibid., p.70). But since Fodor cannot appeal to public criteria for the coherent use of the language of thought, he does need to offer something in their stead.

According to Fodor, the condition for coherent representation in the language of thought will be as follows:

(C*) (S uses 'a is F' to represent *a*'s being *F*) just in case ((S believes that *a* is *F* just in case S is computationally related to 'a is F') is nomologically necessary)

(Here 'a is F' is a formula of the internal code rather than a sentence of a natural language.) The changes in (C*) serve to shift the conditions of coherence from conventional relations, between what people believe and what they say, to lawlike relations between beliefs and computational relations between a subject and his or her internal representations. That is, what Fodor does is use the embedded biconditional to suggest a contingent identity between beliefs, on the one hand, and relations between speakers and their internal formulæ, on the other (this is made explicit at ibid., p.77) and the reason why he must do so is clear.

Unlike the utterance 'a is F' the language of thought formula 'a is F' does not express the belief that *a* is *F*. Unless the formula and the belief are bound together it will be entirely possible that the contents of S's internal formulæ do not match any of S's beliefs, or any of S's expressions of belief. There will, therefore, be no criterion for judging the coherence of of S's representational system and, of course, the system could have no role to play in explanations of what beliefs are.

If, by (C*), S's being in a certain computational relation to the formula, 'a is F', is necessary and sufficient for S's believing that *a* is *F*, and if that, in turn, is necessary and sufficient for 'a is F' meaning that *a* is *F*, then the coherence of the language of thought is as secure as the laws that govern the 'coherence' of the computational relations involved. So long as these laws ensure that 'a is F' is in the right sort of computational relation to S then the formula will be responsible for causing the utterance expressing the belief that *a* is *F*, and S's acting upon the belief. If these relations did not hold then S could not be said to believe that *a* is *F* and, hence, the formula could not, by Fodor's lights, mean *a* is *F*. The computational role of the formula, then, is determined by the laws governing the system's operations. It is this determination which serves as guarantor of the coherence of the language of thought, for it is this that ensures that S means *a* is *F* by 'a is F' if and only if S believes *a* is *F*.

In effect, we are back with the 'physics of the machine' guaranteeing that semantic (in this case coherence) constraints on formulæ are respected by guaranteeing that the sequence of states and operations the machine runs through is such that the terms of the formulæ are used coherently. If, for example, one of the computational relations in question is that of producing the appropriate utterance of the (natural language translation of the) formula, then that relation will, of course, support the assertion that the formula is being used coherently. And, if the formula's production of just that appropriate utterance was determined by the physical construction of the language user then we could agree that the coherence of the entire internal language could be, in principle, secured by that construction. Therefore, the role of linguistic practices or, as Fodor has it, conventions in the public language is fulfilled by causal necessity in the private one.

Of course, the conscription of causal necessity (which, for those who believe all causal relations are covered by laws, implies nomological necessity), as the guarantor of the coherent use of terms, requires that physical laws be construed as having a normative function. But, to use a term coherently is to use it in accordance with norms of use, and these norms, because they can be appealed to in normative evaluation, will be of a kind appropriate to language use and not norms of statistics, of biological functioning, or of mechanical operation. Computational operations, supposedly determined by the physics of the computing machinery, *cannot* explain why the symbols they are thought to range over occur in those operations in a way appropriate to linguistic norms; that is, in a way that can be judged to be coherent. We cannot accept a description of those operations as an explanation of coherence because it makes no sense to speak of the operations as being guided by linguistic norms. Only those uses of symbols that non-accidentally comply with norms are candidates for evaluations such as 'coherent' or 'appropriate' and, as I have maintained, the sort of non-accidental compliance that must be in place in the explanation of what it is to use language correctly—to be competent with a language—is the sort of compliance that attaches to *rule following*. Mechanical operations *are not* instances of rule following because, for one reason, what can follow a rule can make a mistake in the way it does so. But, making a mistake in the following of a rule requires that the rule was being followed deliberately, and machines do not do anything deliberately. Internalizing the rule as a representation is, as I hope to have shown, a futile attempt to account for its guidance of the machine's operations.

The conclusion we should reach, then, is that Fodor cannot provide criteria for the coherence of the language of thought in the form of the physical laws which determine computations ranging over the 'symbols', 'formulæ', or 'representations' of that language. He is, of course, quite right to worry that should there not be criteria he would be wrong to claim that there is an internal language, only the claim is not so much wrong as incoherent. One way of expressing the incoherence is to point out that an explanation of a language's coherence by appeal to physically determined operations and relations involving its symbols, is an explanation that has nothing to do with coherence, for that notion creates the space for criteria which differentiate between sense and nonsense. In public language use the criteria are fulfilled when words are used according to an established practice, although the practice need not be widespread, and might include such prescriptions as that a given word be used, in a given type of context, to refer to an object of a certain type. The prescriptions are rarely a matter of consensus agreement (or convention), as Fodor seems to think, but agreement is at their base. The agreement is displayed by the fact that so much of the time we do not need to define what our words mean in order to communicate with others—we could not communicate if this was always needed—and is best understood as an agreement in judgements as to what words to use and when to use them (cf. Wittgenstein, 1953, §242).

Put another way, the agreement is best understood as in *how* we use our words, such that our judgements concur, not *that* we will use our words so that our judgements concur, for the latter implies that agreement in judgements precede the agreement in use of words whilst the former allows that the agreement in judgement just is the agreement in how we use words.[7] The distinction I make is important but it need not be laboured over here for neither route to an explanation of linguistic coherence is available to anyone who supposes that there is an internal language. The feature of internal language use which makes these explanations unavailable is its inability to establish linguistic norms. For even if we can speak of agreements in the use of the internal language, so that the agreement is between uses of the internal language by one organism, over a period of time, or between uses by many organisms the mechanistic norms which are established will not give rise to normative evaluation. That is, a mechanical regularity does not set a standard by which to judge a machine's functioning as correct. After all, one's watch may always gain time, but if after hitting it on the table, it keeps good time, it is not incorrect.

The inapplicability of normative evaluation is owed to the fact that mechanistic norms, whether they be functional descriptions of particular machines or the laws of physics, are not followed by machines, organic or inorganic. Machines do not learn to apply them, make

[7] We might note that taking the agreement to be in *how* we use words discourages the supposition that we can talk about the use of words *in vitro*, so to speak. That is, it sits awkwardly with the thesis that to use words correctly is simply to construct well-formed sentences, irrespective of context—a view one might have if one took language to be a calculus of syntactic rules.

mistakes in doing so, and correct their mistakes in the light of them. It makes no sense to appeal to these mechanistic norms, therefore, in explaining what it is for a machine's functioning to be coherent, and so, if this is all that can be appealed to in accounting for the coherence of a machine language then it makes no sense to talk of coherence in such a language. Of course, as Fodor realized, if there is no coherence in the use of a language then we are no longer talking of a language.

3 CONNECTIONISM

The conclusion we have reached is that there is no such thing as an internal language. This means that the explanation of linguistic competence, and cognition, as dependent upon the rule-governed manipulation of internal symbols is not a viable enterprise for a cognitive science. However, it is not just the Classical Computationalist approach of Fodor which must relinquish its claim to cogency. Those, like Stich, who have attempted to avoid the problems of attributing content to internal items by postulating a purely syntactic internal language, will also have to forego the claim.

Although the syntactic model, proposed by Stich, required a grammar for the generation of complex syntactic objects from a finite set of primitives (see chapter 2 above), his combinatorial model does not rely on context independent *meaning* atoms and so evades the objections faced by Fodor. However, in requiring a grammar and in identifying the syntactic complexes with *well-formed formulæ* he becomes a target for the arguments against rule following mechanisms. Ultimately he too faces the charge of incoherence levelled against anyone who posits an internal language.

Stich jettisoned his Syntactic Theory of Mind in favour of the Connectionist approach to cognitive explanation and, in doing so, added a new justification for the purgation, of folk psychological explanation from cognitive science, called for in his earlier work. While I believe the Eliminativist conclusion he pressed for is entirely misconceived, the argument he produced (in collaboration with Ramsey and Garon), against attempts to place propositional attitude states within Connectionist networks, was well directed. If beliefs are states of the brain with propositional content, then those states ought to display propositional modularity and, we might add, if propositions are represented in the brain, then there must be a language in which the brain expresses them. Thus, the Connectionist who wishes to uphold propositional attitude Realism needs a system—to all intents and purposes, a symbolic language—as a medium in which to represent the propositional contents of belief states. However, apart from facing the tortuous task of specifying the nature of representation in a Connectionist system, a task which appeared to be fulfilled only at the cost of content epiphenomenalism, the Connectionist also faces the body of objections which stands in the way of representationalism of any kind. The representations of Connectionist networks cannot be representations *for* those networks unless they are used *as* representations by them. If the networks really did *use* representations then the cognitive faculties that they are supposed to explain, such as memory, concept learning, and pattern recognition, would be duplicated within the networks, thereby generating a vicious circularity and regress; a consequence supporting the conclusion that the proposed explanations are vacuous. The option of claiming that a network *uses* representations in the sense that they have a causal role not only seems unavailable, given the causal inefficacy of Connectionist representations, but also requires an identification of normative and causal role which we have ruled out. But without representations, and representational content, the networks will not support intentional states, for the idea of an intentional state is the idea of a state with content; that is, a state individuated by what it is about.

Of course, if Smolensky had provided an account of the constituency of representational content, to the satisfaction of Fodor, Pylyshyn, and McLaughlin, that account would have fallen squarely into the target area of the arguments I offered in chapter 5 against detachable meaning constituents. In other words the flaw in his theorizing is not his failure to satisfy his critics by providing an account of the productivity and systematicity of thought and language, rather it is his attempt to isolate, in the form of Tensor Product Representations, units of meaning.

The meaning of a word, or any symbol, cannot be seen as belonging to it such that it is transported into any context in which the word is placed. If it could then we should not be able use the same words in both denotative and attributive roles, and we can. We have not explored the possibility of a Connectionist account of mental content and linguistic competence deeply enough to warrant the judgement that they cannot reply to such an objection. However, let us speculate upon what such an account might be.

Following the contours of Smolensky's thinking we might say that a network can be 'trained' such that it will represent a concept (such as *coffee*) as a collection of microfeatures (like *hot liquid* and *bitter tasting*) defining that concept, and such that it will respond to their instantiation in the environment by expressing the concept in the sentential output of the network. We might speculate further that whether what is symbolized in the output is the concept or an instantiation of the concept (an object) will depend upon the position of the symbol in the output sentence.

Laying to one side misgivings about the notion of representation and the Classicist's charge of content epiphenomenalism, we should comment that, although the movement away from compositional semantics *should* be an improvement, without compositionality the Connectionist system will have no apparent way of accounting for misrepresentation. For unless a representation has a propositional content (or unless it occurs in a context in which it is understood as asserting something, as in the case of a one word answer to the right sort of question) it cannot be true or false. This may be acceptable for those

Connectionists who are uncomfortable with propositional attitude Realism but then it would no longer be clear just what it is they take themselves to be modelling in their networks—they could say nothing about thought *processes* or belief *states*, for example. Supposing, then, that propositional attitudes are not among the explananda of a Connectionist theory of cognition, perhaps the concern is with linguistic competence alone.

The objection Connectionism faces is that without an account of what it is for a network to believe, intend, understand, and so on, it will not be possible to attribute it with linguistic competence. Such competence can be attributed only to what can, among many things, use words correctly or incorrectly; make mistakes in, or misconstrue instruction on, word use; or choose the right or wrong words to express itself. A system's production of sounds or inscriptions, deemed appropriate to a context, does not amount to linguistic competence. The propriety of that production may be simply fortuitous, or it may be the result of careful programming of the system, but in neither case will it be indicative of the system's competence. In attributing competence to a speaker we are offering a judgement based on the evaluation of what is said or done. Unless we can take for granted that what is said and done is *intended* to express beliefs, show understanding, or to indicate facts, for example, we are not in a position to judge whether the speaker's use of words is correct or incorrect, well chosen or misjudged. So if a system, or network, cannot be said to intend, believe, or understand, anything then it cannot be competent in language use.

Note that the use of the word 'training' cannot be of help here. The training set to which the network is exposed, in implanting a 'concept' within it, does not constitute a norm of language use. If the network's yields, as output, an application of a word we might deem to be sub-standard it might be due to the incompleteness of the training procedure—the backward propagation algorithm may yet to have stabilized the network's response to inputs—or to something like a chemical imbalance within the neuronal cells, but neither case amounts to the making of an error or mistake, nor to incompetence. Insofar as the training amounts to a programming of the network we can say that the output is indeed incorrect in the light of a linguistic norm, but if there is any incompetence involved, it is the programmer's and not the network's. The important point is that unless the training can be seen as leading to an understanding of a concept, that is, of the use of a word, then there will never be a place for the notion of competence in describing what the network does. When training is nothing more than programming the place is lacking and the possibility of any linguistic competence is removed.

Clark's more recent brand of Connectionism, briefly examined at the end of chapter 3, would appear more promising in this respect. In eschewing an ontology of 'folk solids', such as concepts and intentional states, he is able to recast linguistic competence as the possession of abilities to behave in ways indicative both of intending and understanding the use of words. Roughly speaking, for Clark understanding a word is a matter of being able to respond appropriately to its use, and being able to use it appropriately. However, any plausibility in Clark's account of concept possession soon evaporates when we discover that in saying that a person grasps a concept 'we are really ascribing a *body of knowledge* and skills whose manifestations may be both *internally* disparate (in terms of occurent *representational states*) and externally disparate...' (Clark, 1993, p.204, my italics). I should like to comment on the line of thought intimated by this brief quotation.

The suggestion that understanding and using words, or a language, requires a *body of knowledge* is the source of, what is sometimes called, the problem of linguistic creativity. The problem, to which the combinatorial story of Chomsky, and later Fodor, was proposed as a solution, arises from the question of how we are able to construct and understand an unbounded set of sentences on the basis of an exposure to a, relatively, small number of exemplars. However, this becomes a problem only when the ability to speak a language is understood as requiring an explanation of what things we must *know* in order to do so. When we understand the ability in this way we are naturally tempted to seek solutions by positing innate languages, or, at the very least, looking for some means of storing *internally* a knowledge of how to use a language; that is, we are tempted to posit internal *representational states*. The temptation is difficult to resist because we are thinking of linguistic competence as a matter of knowing certain principles and definitions which are put into practice as we engage in linguistic activity.

Much of the argumentation I have offered points to the incoherence of such a way of thinking about our ability to use a language. If that ability did require a body of knowledge, if, that is, one needed to learn certain facts about the meanings of words and the rules governing their use before one could use them, then the likes of Chomsky and Fodor could be forgiven for claiming that there must be an innate language of thought in which to represent these facts and rules. But if, as I have argued, the claim is incoherent then the linguistic ability does not require a body of knowledge.

That said, it would be wrong to suppose that the body of knowledge Clark has in mind (so to speak) is linguistic. He suggests that it is manifested, rather than carried, in representational states. That is, he is not explicitly committed to knowledge of a language being knowledge *that*, for example, the use of a certain word is governed by a certain rule. That is, it is open to Clark to say that knowledge of how to use a language is *tacit* knowledge. Of course, if knowledge of a language was tacit, then talk of 'a body of knowledge' would be out of place. Tacit knowledge, insofar as one can have it, is not possessed as a body of things that one knows.

Clark's 'body of knowledge' *is* a body of stored knowledge, only it is not stored in the form of propositions to be retrieved for the purposes of processing, as the Classical Computationalist would have it. Rather it is stored as connection weights, for 'these weights just *are* the network's store of knowledge' (ibid., p.39), and they are therefore intrinsic to the network's processing rather than extrinsic, or related by retrieval, to it. This gives us the sense in which the body of knowledge is 'manifested' in occurent representational states since it is the patterns of activation, created by 'use' of connection weights that constitute those representational states (ibid.). If this is the body of knowledge entailed by the grasping of concepts then what we are being asked to accept is that using one's knowledge of languageis not simply a matter of, for example, using words to meet certain ends—to relate facts to others, to make requests, to give commands, to offer encouragement, to tell jokes, and so on—but, rather, using the knowledge is to be identified with the processing occurring within one's neural networks.

This amounts to a skewed view of what it is to use one's knowledge of language for, after all, one *can* misuse one's knowledge of how to use words (by inciting a mob to riot, for example) but that misuse cannot be seen as a neural process. Such facts could be missed only if one shares the conviction, common in cognitive science, that we cannot say that knowledge of how to speak a language, or of how to use words, is simply a matter of being able to do so. In cognitive science the conviction gives rise to the insistence that what is needed is an ætiology of linguistic competence; that is, what we require is an account of what it is *in* human beings that causes the manifestations of linguistic competence; of what causes linguistic performance. The question of how we are able to formulate and understand an infinite number of sentences seems to demand an answer in terms of what we must know when we know how to use language and, as the defining problem of a scientific enterprise, this answer requires that the knowledge be understood as a 'body' of some sort, to be found within, which can be identified as the cause of linguistic performance. The answer which must be rejected, on this understanding of the problem, is that one is able to produce and understand novel sentences simply because one has learned, and so knows how, to speak a language. Any vacuity detected in the answer is to be expected since what it provides is a seldom required criterion for the correctness of saying that someone has learned, is able, or knows how, to speak a language. Someone unable to produce and understand novel sentences would not have learned a language. But, any vacuity in the answer is merely a reflection of the vacuity of the question.[8]

The dissatisfaction that might greet this answer is to be met by the reply that answers in terms of internal rule-governed representational systems just are not answers. The incoherence of the notions of inner representation and internal languages prevents them from being answers. Clark's attempt at explaining concept possession, though apparently free from the trappings of compositional accounts of language, reflects the common conviction that unless there is a systematic answer to the question of linguistic creativity it will forever remain a mystery how anyone can speak a language. The existence of such a conviction was known to Wittgenstein whose antidote to it was as follows:

> 'if I talk or write there is, I assume, a system of impulses going out from my brain and correlated with my spoken or written thoughts. But why should the *system* continue further in the direction of the centre? Why should this order not proceed, so to speak, out of chaos? The case would be like the following—certain kinds of plants multiply by seed, so that a seed always produces a plant of the same kind as that from which it was produced—but *nothing* in the seed corresponds to the plant which comes from it; so that it is impossible to infer the properties or structure of the plant from those of the seed that it comes out of—this can only be done from the *history* of the seed. So an organism might come into being even out of something quite amorphous, as it were causelessly; and there is no reason why this should not really hold for our thoughts, and hence for our talking and writing' (Wittgenstein, 1981, §608).

[8] Eugen Fischer has offered a related argument to the effect that the problem of linguistic creativity is to be dissolved rather than resolved. His argument is directed at those who would see the problem as that of how one comes to understand the meaning of a sentence on the basis of an understanding of the meaning of its parts. This way of seeing the problem suggests the need of a solution in the form of a specification of a semantic theory possessed by speakers of a language. Roughly, Fischer's response is to question this need by arguing that we cannot make sense of the assumption that one could understand the words of a sentence without understanding the sentence itself (Fischer, 1997).

The requirement that linguistic competence, and cognition, be explained in terms of inner processing is the requirement that we find a material *cause* for our uses of language and our cognitive capacities. It is a consequence of the conviction that such phenomena *must* have a causal explanation; that only such an explanation really tells us how we are able to speak, understand, and think. The denial of such explanation seems to lead to a loss of the power to predict, and generalize about, human behaviour. However, this is not the case. In Wittgenstein's analogy, we have a means of predicting, and generalizing about, what the seeds will grow into because we have the history of the seeds to go on. We know that A-plants develop from A-plant seeds and B-plants develop from B-plant seeds. And;

> 'If I say: the history can't be the cause of the development, then this doesn't mean that I can't predict the development from the previous history, since that's what I do. It means rather that we don't call *that* a 'causal connection', that this isn't a case of predicting the effect from the cause.
>
> And to protest: "There *must* be a difference in the seeds, even if we don't discover it", doesn't alter the facts, it only shows what a powerful urge we have to see everything in terms of cause and effect' (Wittgenstein, 1976, p.375)

In the case of our linguistic and cognitive abilities what allows us to explain how people can produce an indefinite number of novel sentences or entertain any number of thoughts is that they have learned a language. That is, they have grown up in the society of adults who taught them how to ask for things, to fetch and make things, to tell when they are hungry, tired, or bored, and who read them stories, encouraged them to make up their own, and so on. When we know a child has learned a language we can talk to him and expect to be understood (although we must make allowances for his age), and we can predict that he will be able to solve certain problems (like how to get the biscuit tin down from the shelf where mother thought she had hidden it). We can explain and predict things about the child because we know something about his history and, to be sure, the explanations we arrive at are not 'causal' (in the sense of that word as we find it used in the sciences). But that does not mean we should look for the causes inside the child and think of learning as a process hidden from us; for the phenomena of learning are not hidden. The words we use in connection with language learning includes 'memorizing', 'reciting', 'copying', 'experimenting', 'questioning', and 'listening', and these are words describing what *children* do, not their *brains* or cognitive *mechanisms*.

It is worth noting that Connectionists often believe that they have produced an account of learning that improves upon the Classical Computationalist model in allowing that neural networks are 'trained' by being presented with inputs and by 'correcting' their outputs through employment of a backward propagation algorithm. They also believe that they have an account of concepts for Connectionist networks which allows for the fact that there are often no features common to all instantiations of a concept word—the point Wittgenstein makes when he talks of 'family resemblance' in connection with the use of the words 'language', 'number' and 'game' (Wittgenstein, 1953, §§65ff). The idea might be that provided an input activates the network such that it will acquire a vectorial value within the vector space which may be said to represent a given concept (to put it in Smolensky's terms) then the network will apply that concept to the input stimulus. The input values need not have any essential property in common. Moreover, the plasticity of the weightings and biases of the network allows for conceptual revision and for a variation between the occurrent inputs and the training inputs in a way that explains the novel application of concepts required to explain linguistic creativity.

The salient point about the conjunction of these accounts (of learning concepts, on the one hand, and of the indefinitude of the concepts learned, on the other) is their incompatibility. 'Training' and 'learning' are thought to be explicable as particular processes in which networks are presented examples, in the form of inputs, in order to preset activation patterns such that a stimulus, as input, will generate a correct output. The illicit importation of the notion of correctness is, of course, worthy of comment—neural networks are no more capable of correct or incorrect behaviour than more traditional cognitive mechanisms—but it is the incompatibility of the accounts to which I wish to draw attention. I shall do so by quoting from *Computers, Minds and Conduct*, by Button, Coulter, Lee, and Sharrock who observe;

> 'There is no *uniform* phenomenon called 'learning from examples' (eg., being presented with many cases of the same kind of phenomenon until it can be correctly reproduced or simply named). And we must remember that what is being trained is not a brain but a human being; what a human being is being trained to do under the rubric of 'learning what a word means' (even a referentially usable one) is far more than correctly to repeat it in the presence of an instance of the named phenomenon; and what is doing the 'learning' is a person, not his brain. Conflating the technical concept of 'training' (as in 'training the net') with its non-technical, ordinary and human-level applications, as in, for example, 'training a marine recruit', 'training a child in hygienic habits', or training an infant in the multiplication table', results, not in better scientific theories, but in rampant confusion' (Button, Coulter, Lee, and Sharrock, 1995, pp.130-131).

There is nothing wrong, therefore, in using the word 'training' to refer to the manner in which the connection weights and biases of a Connectionist network are modified by input sets and backward propagation, nor with extending the use to cover what happens in neural processing, but there is a lot wrong with taking this use of 'training' to be essentially the same as the ordinary use; that is, with taking this use to refer to what is essential to the concept of 'training' as it is applied to people. It should be clear that talk of training brains cannot be conflated with talk of training recruits, children, and infants, for in the course of teaching a person one might encourage, rebuke, cajole, praise, demoralize, build confidence, and, of course, correct mistakes and errors. It makes no sense to talk of training a neural network in this way. Indeed, if, as I argued in the previous chapter, there is no sense in talking of mechanisms *following* rules, and training in the use of words requires that one learns to follow rules, then it follows that it is nonsense to speak of brains, or parts thereof, being trained in the use of words. The Connectionist's usage of 'training' and 'learning' might be a helpful shorthand when describing the modifications of parallel distributed processing networks, with which they are concerned when (innocently) working on projects in cybernetics, but it must be remembered that this use is very different from those the words receive when applied to human activity. As Button et al put it, 'this new usage would not license any claims on behalf of a *psychological* "learning theory" pertinent to human language acquisition (or anything else)' (ibid.).

The discussion of this section should suggest that a Connectionist science of cognition is only a marginally more intelligible prospect than a Classical Computationalist or Functionalist science. Connectionists often make allowances for such things as the context dependency of symbol meaning, the indefinite nature of many concepts, and the acquisition of concepts as akin to the acquisition of abilities. In doing so they improve upon the Classical modelling of cognition and linguistic competence and, distressingly, these improvements are largely due to the influence of Wittgenstein's thinking about language on their theorizing. Indeed, this influence is sometimes made explicit (see Goldstein and Slater, 1998, and Mills, 1993)[9]. It is distressing that Wittgenstein can be so misunderstood, for if these theorists really did adopt an approach to thinking about language and cognition in line with Wittgenstein then, I would suggest, they would cease to be cognitive scientists. They would no longer see the need to account for linguistic understanding in terms of internal representations, and they would refrain from talking of cognition in terms of processes occurring in the brain. To paraphrase and then summarize Wittgenstein on this matter, it is

[9] The following quotation from Goldstein and Slater exemplifies the propensity of some theorists to cite Wittgenstein in support of decidedly un-Wittgensteinian theses:

> 'We have argued that if the neural state, or any part of it, were a representation, it would be idling. It would just "hang in the air" (*PI* §198), itself standing in need of interpretation (*PI* § 210). The word "Mama" represents the mother, but that symbol is external to the brain, *and all the child's brain learns to do is use the symbol appropriately*, in an equally external context.' (Goldstein and Slater, 1998, p.313, my italics).

no more essential to the understanding of a proposition that one should have an internal representation in connection with it, than that one should make a sketch from it (1953, §396) and the concept of 'thinking'—our ordinary use of that word—simply does not fit with the supposition that it is a process located in the brain, or anywhere else for that matter (ibid. §§359-361).

The objections I have raised against the notion of internal representational (or propositional) content, and against the very notion of internal representation, suggest that Connectionists could not lend support to propositional attitude Realism even if they wanted to, but they also suggest that the explanation of linguistic competence in terms of internal states and processes is equally insupportable. These objections, like the ones I raised against a causal (and, in this sense, naturalistic) account of normativity, have a broadly Wittgensteinian provenance[10] and, if they are objections that should be sustained then two things will follow. Firstly, Connectionists who believe that their approach to explaining cognition and linguistic ability is compatible with Wittgenstein's thinking are mistaken. Secondly, the shortcomings of their approach, insofar as it represents the best form of explanation that cognitive science can offer, indicate that cognitive science should be deemed moribund.

[10] I attributed the context principle to Frege when, in chapter 5, I cited it in arguing against the assumption of context independent meaning. Its importance was clearly recognized by Wittgenstein, for he formulated it in the *Tractatus* (Wittgenstein, 1922, §3.3) and adhered to it in the *Philosophical Investigations* (see, for example, §§39-43). Indeed, the importance of the principle to Wittgenstein is proportionate to the prominence, within the *Investigations*, of the insight that, for a large class of cases, the meaning of a word is its use (§43).

CHAPTER EIGHT

EXPLAINING AWAY COGNITION

In the last four chapters I have tried to demonstrate that Realism about propositional attitudes is incoherent. I would suggest that any Realist about mental states needs to account for their content and, like Fodor, I would suggest that the Representational Theory of Mind offers the clearest means to this end. Realists who decline the offer and prefer to think of beliefs as monadic functional states must admit that such functional states are not states unless it they are instantiated and, as soon as they do, they must explain how those state are about something. It is difficult to see how that can be done without commandeering the concept of 'representation'. Despite the differences both versions of Realism make the claims, firstly, that verbs like 'believe', 'intend', 'understand', 'remember', and 'think', *refer* to states, events, or processes, occurring within a subject and, secondly, that these occurrences are *causally related* to one another, to external occurrences and objects, and to behaviour.

If the arguments I have offered are valid they will suffice to show that Realism is untenable and, in doing so, undermine a good deal of the theorizing pursued by cognitive scientists. However, one possible conclusion to draw from this is that to reject Realism is to accept of Eliminative Materialism. My aim in what follows is show that this is a false alternative by demonstrating that the reasoning which yields the conclusion is erroneous. Let us begin with a brief rehearsal of that reasoning.

As we saw in chapter 2 there are degrees of Eliminativism ranging from the claim that the psychological vocabulary should be removed from general discourse, to the suggestion that it should be gradually removed from scientific accounts of the ætiology of human behaviour. However, the feature common to all whose views fall within this range is an acceptance of the premiss that the psychological vocabulary is embedded in a theoretical framework, referred to as 'folk psychology'. When this premiss is augmented by another, to the effect that the folk theory is false, the conclusion follows that the theory should be displaced. The position of the Eliminativist within the range is, therefore, largely determined by his or her convictions as to what the theory should be displaced from, and when it should be displaced. Thus Stich suggests that folk psychology, since its role is not purely theoretical, should be removed only from the scientific explanation of cognition (although he sometimes makes bolder claims), whilst Patricia Churchland suggests that psychological terms can be used as a way of delineating the explanatory domain of neuropsychology until such time as it develops its own delineation of that domain. Of course, since the neuroscientific explanation can be expected to cross-classify the folk psychological one, we can expect the latter to become obsolete even in general discourse. The important point is that, however Eliminativists might differ in their prescriptions for treating folk psychology, they all conduct their diagnosis of its explanatory maladies on the premiss that it is a theory.

1 THE FOLK THEORY

The claim that folk psychology is a theory is grounded upon two assumptions: Firstly, it is assumed that the psychological vocabulary is straightforwardly referential. In this respect folk psychology will entail the first claim of propositional attitude Realism as well as Realism about sensations and perceptions. (As our concern is with Realism as it pertains to cognition I will have little to say, except by implication, about the other forms). Secondly, it is assumed that employment of the vocabulary entails, at least tacit, acceptance of a body of generalizations about the relations psychological entities enter with other psychological entities, external objects, and behaviour. This is not quite the same as the second claim of

Realism, the causal thesis, but that thesis is implied by the first claim (and the Eliminativist's first assumption) together with acceptance of the view that singular causal relations fall under laws. Working backwards we can reason as follows: If psychological states enter into relations over which generalizations can be made then the type of relations these will be depends upon the nature of the relata. Since some of these relata are physical objects and events about which we have beliefs and desires, on the one hand, and bodily movements which result from those beliefs and desires, on the other, it seems fair to say that the folk, who are committed to the generalizations, will presume that the psychological vocabulary refers to states having causes and effects. Thus, the generalizations the folk are assumed to be, albeit tacitly, acquainted with will be generalizations about causal relations. The claim that folk psychology is a theory, then, amounts to the claim that those who adopt it are Realists in the sense outlined above.

Since folk psychology is to include the 'common-sense', everyday, use of psychological terms by the general population, the claim that it is a theory amounts to the claim that by using these terms we give tacit consent to these two assumptions or, in other words, if these assumptions were made explicit to us we would be bound either to consent to them or to refrain from using the terms. Bearing this in mind we can rule out some further assumptions, about the theoretical commitments of folk psychology, made by all three of the Eliminativists we have cited.

Several of the arguments offered in favour of rejecting the folk theory are directed against various forms of the Representational Theory of Mind. The last of the six arguments of Paul Churchland (see chapter 2) is directed against the Ideal Sentential Automaton approach which posits sentential representations by way of accounting for propositional attitudes. The argument is revised as the 'Infralinguistic Catastrophe', by Patricia Churchland and Stich, as a means to undermining sentential attitude psychology and both add further arguments (based on the problem of tacit knowledge, the frame problem, and the supposed commitment of folk psychology to propositional modularity) to justify their brands of Eliminativism.

Now, while I am ready to agree that these arguments pose real problems for representationalists, I think it is simply absurd to suppose that use of psychological expressions commits one to any form of representationalism. Is it reasonable to suppose that anyone would respond to the remark 'I thought you had left' with 'You mean you internally represented my leaving'? The suggestion that the second sentence makes explicit a tacit assumption we all make would be intelligible only if the assumption acted as a suppressed premiss in any reasoning we may ordinarily be led to on hearing the first sentence. However, the supposition that in uttering the first sentence someone has reported on the presence of an internal representation, the content of which he or she believes, is extraordinary and should not be attributed to anyone with common sense. Furthermore, it would be absurd to suggest that on being persuaded of the fact that there are no such internal representations, we would have to admit that we do not think, believe, remember, and so on.

That said, one can see why, if it is thought that folk psychology is committed to propositional attitude Realism, it may be supposed a tacit commitment to representationalism lurks in the background. As I indicated in chapter 4, the claim that beliefs are internal states seems to require an account of their content in terms of representations. However, if one acknowledges that there is no such tacit commitment then the natural conclusion to draw is not that the folk Realism is inconsistent, but that the folk are not committed to Realism. Moreover, since the two assumptions made by Eliminativists, in support of the premiss that folk psychology is theoretical, coincide with the two claims of Realism, it follows that in as much as it is not a Realism, folk psychology is not a theory. However, the entailment from Realism to representationalism is, perhaps, not strict enough to warrant the conclusion that the absurdity of attributing the latter to the folk implies the absurdity of attributing to them the former. In what follows, then, I will offer some reasons to reject each of the two Eliminativist assumptions. It would require more space than is available here to give these reasons in full but I am inclined to believe that the outline I will give, when combined with the consideration with which I close this thesis, support the view that the Eliminativist conclusion is the misbegotten offspring of two misconceived premisses.

2 THE REFERENCE OF PSYCHOLOGICAL TERMS

It would be wrong to suggest that we cannot refer to beliefs, thoughts, memories, and the like. The question 'Do you remember that you once believed electricity is a liquid?', when asked of a particular person, would be a case of referring to a past belief. However, this sense of 'refer' differs from that used in the context of an explanation of a scientific taxonomy where terms of a vocabulary are given a reference to types of state, event, process, force, relation, and so on. The difference is hinted at by the strangeness of formulating the question as 'Do you remember that you once were in a state of believing that electricity is a liquid?'. However, this difference is not recognized by those who would view psychological terms as theoretical.

To understand the sense in which psychological terms are presumed to refer by Eliminativists we might recall an example, offered by both Patricia and Paul Churchland, of a theory which has been eliminated and which serves as a precedent for the elimination of folk psychology. Phlogiston theory posited a substance (phlogiston) inhering in matter susceptible to combustion and corrosion, and released when those processes occur. However, the development of pneumatic chemistry led to the denial of the existence of phlogiston and the positing of another substance (oxygen) which explained the phenomena for which phlogiston was cited as a cause. Phlogiston could not be reduced to oxygen, for although phlogiston was, and oxygen is, seen as a requisite for combustion, oxygen is thought to be drawn into the process from the atmosphere whereas phlogiston was supposedly present in the burning matter. In effect, phlogiston theory posited a substance which the more successful oxygen theory did not recognize in its taxonomy. So although 'phlogiston' was taken to refer to a type of substance, it was found that there was no such substance. Now, if the relation between the two theories is to be analogous to the relation between folk psychology and a mature neuroscience, then psychological terms should refer in a comparable way to both a term like 'phlogiston' and terms used in neuroscience.

The entities referred to in the psychological vocabulary will be states (of belief, for example), events (like thoughts), and processes (such as deliberating).[1] Thus, when we ascribe beliefs, thoughts, and deliberations to ourselves and others we are ascribing states, events, and processes. (To avoid needless repetition I will restrict the discussion to talk about states). It should then follow that the criteria upon which we base these ascriptions are criteria for identifying states. Of course, it need not follow that these criteria are directly observational (for the identification of a state as gaseous need not be based on direct observation of a gas) but, when they are not, they will be indirectly observational in that the presence of the state is inferred from what is observed directly in virtue of its causal properties. Criteria for identifying a gaseous state, for example, include the detection of molecules of a substance, in the atmosphere, by an instrument designed for the purpose.

Now, since we are considering cognitive *states* it will be helpful to get a general idea of what sort of criteria we use for individuating states. Water, for example, can be found in a suspended, liquid, and solid state depending upon the temperature and pressure to which it is subjected. A classroom experiment to discover the temperature at which water changes from

[1] It will not matter to the argument whether one wishes to say, for example, that the only ontological category we should recognise, within the temporal dimension, is that of events.

solid to liquid state at normal atmospheric pressure might involve observing a block of ice, waiting for it to begin melting, and then collecting and measuring the temperature of the melt-water. The pupil will need to monitor the ice and watch for drips which she can collect and measure. In doing so she will be monitoring the change of state from solid to liquid form. We can imagine her being asked, periodically, by the teacher 'Has it started melting yet?' and her calling out 'It's melting now' or perhaps 'It's turning into water now'. Another experiment might involve timing how long it takes for a quantity of ice to melt completely in a enclosed environment kept at a constant temperature. We would expect similar questions and answers—'Has it all melted yet?', for example. Thus, there are observational criteria for distinguishing states of a substance and for determining the onset and cessation of different states.

The rudimentary example will differ in complexity of observational procedure, but not in principle, from individuation of states of the brain where activity states of various of its regions might be measured using, for example, Positron-Emission Tomography (PET scanning). PET scanning could, in theory, allow one to monitor the duration of an activity state in, say, the hippocampus, within certain pre-specified parameters of intensity and required percentage of the region in that state. So similar questions, as to the onset and cessation of the state, asked of the pupil, could be asked of the neuroscientist. We might add that these considerations about the criteria for individuating states and state changes in physical media apply also to the individuation of events and processes, where observations and measurements will allow the duration and frequency of both to be monitored.

Now, if terms of the folk psychological vocabulary are used, by the folk, to refer to states in the way that terms like 'solid', 'liquid', and 'activity' are used to refer to them, then we should expect the folk to employ similar criteria for the individuation of psychological states. We may consider these criteria from the first and third person perspective which correspond to the perspectives of the pupil and the teacher, or the neuroscientist reading the results of the PET scanning and an interested colleague.

Beginning with the third person perspective, someone, let us call her Pat, might observe a flatmate Paul (who has lived, until recently, in an Amish community), putting containers under the electrical socket outlets in the kitchen. She asks why he is doing this and he tells her 'It's in case any electricity drips out'. Pat, in surprise, asks 'What! You mean you believe electricity is a liquid?' and is answered in the affirmative. According to the Eliminativist assumption, in asking the question Pat has enquired as to the existence of a state of Paul's mind we might call 'the belief that electricity is a liquid'. Paul's affirmation confirms the existence of the state. If this is the case, and if this is a state in the same, or similar, sense in which liquidity or neural activity are states, then it will follow that certain sorts of question, like, 'When did you begin believing that?' would be appropriate ones for Pat to ask. Such a question might be asked, although 'Who told you that?' would both be more natural and reflect what Pat would want to know—for it would not matter to her if there was a precise time Paul began believing. But would it be appropriate for Pat to ask Paul, periodically and prior to correcting him, 'Do you still believe that electricity is a liquid?' or 'Are you still in that state of belief?', or to say 'When I explain to you what electricity is, I want you to say "now" at the point at which you cease to believe it is a liquid'? The answer is that these would not just be inappropriate but outlandish things for Pat to ask and say. This suggests that individuating such a belief is not individuating a state in anything like the way that states are individuated in theoretical investigations. The criteria for third person ascriptions of such beliefs are not criteria for the individuation of states of mind.

Of course, other psychological terms might be ascribed on the basis of observation of states. A state of anger or fear, can have what Wittgenstein calls 'genuine duration' (see

Wittgenstein, 1981, §§45 and 81-85)[2] because, as with physical states, it would be possible to ask an observer to attend to the state and indicate when it alters or subsides. But what is observed is the state of the *person*, not his mind or central nervous system, manifested by certain behavioural symptoms, such as agitated gestures, facial expressions, and physiological symptoms such as a flushed complexion, tensing of muscles, and perspiration.

However, two things should be noted. Firstly, these symptoms alone are not sufficient for the individuation of a state as one of anger or fear, for the same (non-verbal) behavioural and physiological symptoms could occur—in that they might overlap—in cases of both anger and fear, and even of awe, joy, and excited anticipation. What allows us to distinguish between correct and incorrect ascription is the context of the behaviour and physiology so that, for example, the agitated gestures of a person who has been insulted by a another indicate that he is angry. Hence, the truth of a statement like 'Smith is in a rage' cannot be judged solely on observation of Smith's behaviour whereas, 'The water is in solid state' can be verified solely by observation of the sample. In the case of observations of states of substances contextual features do not have a state individuating role and, since this suggests that the criteria for correctly individuating the states of anger and fear are not those for individuating states of a substance, it would be wrong to say that they are 'states' in the same sense.

Secondly, the parallel between first and third person observation, of both states of mind and of a substance, is lost when we *do* identify a state of anger of fear as having genuine duration. That is, the observer of the angry person, rather than the person himself, is the one who is able to observe the duration. The significance of this fact is that it leads to another respect in which applying psychological terms is not comparable to individuating states of a

[2] For an exemplary application of the concept of 'genuine duration' to the question of whether mental terms refer to states of mind see Norman Malcolm's criticism of the Causal Theory of Mind in Armstrong and Malcolm, 1984, pp.79-86, 201&202.

substance; first person psychological ascriptions are not based on observations of states.

When Paul tells Pat what he believes, or when the arachniphobe tells us that he is afraid, it is not on the basis of an observation of his state of belief or fear. There is much to speak in favour of this denial. In the case of Paul's state of belief, if it had genuine duration then it would make sense for him to ascertain for himself, from time to time, whether he still had that belief and, clearly, that is absurd. The existence of some beliefs can be determined in this way, such as a belief in the sanctity of marriage, perhaps, but the determination is not based on the contemplation of a state but of what one might be inclined to say or think about marriage. Similarly, it would not make sense for Paul to wonder whether he still believes that electricity is a liquid while he is asleep, but he might wonder if a substance will change state in that time.

We might also consider what it would be like for Paul to identify his state of belief. Given that he does not do so on the basis of observed behaviour it would seem that he must detect some property of beliefs introspectively. That is, if we are to understand statements of the form 'I believe that *p*' as based upon the subject's identification of a belief state, it must be that there is something that the subject, at least, purports to identify internally. Of course, apart from the implausibility of such a requirement—we cannot seriously suppose that Paul might answer the question 'Do you still believe electricity is a liquid?' with 'Wait a moment, I'll just check the contents of my mind'—the implication would be that the first person use of the word 'belief' is taken, by the folk, to be the result of connecting the word to a private referent. But then, if, on another occasion on which Paul is questioned as to why he is placing containers under the electrical sockets, there had been a change in his internal state (so that the state referred to by the original belief ascription was no longer existent), would we then say he would be wrong to self-ascribe the same belief? And, if so then what criteria would he, or we, have for saying that it was not the same state? Since the answer to the second of these questions is 'None', the first cannot be asked intelligibly. If that question is unintelligible, then the claim that we use words like belief to name internal states cannot be upheld.

Although an arachniphobe's state of fear may have genuine duration the avowal 'I'm afraid of that spider' is not a report. on the existence of the state, based upon observation of behaviour and physiology. To be sure. a fear might be brought to one's attention when one puzzles at one's reaction to an object, person, or occurrence, but the reaction is not enough to generate the conclusion that one is afraid—in another situation the same reaction could indicate awe, anger, excited expectation, and so on. Thus, it is one's reaction *in a situation* which would lead to the realization that one actually fears something, and that is not something one observes in anything like the way one might observe the state of a substance or object.

Then again, the self-ascription of fear is not based upon observation of an internal state, whether that be a state of the central nervous system or a state individuated by introspection. Clearly, ordinary self-ascriptions cannot be reports on the states of our central nervous systems unless it is argued that reports based on introspection are intended to be reports on such states. But to argue in this way is to maintain both that the folk explicitly assent to a Central State Identity Theory, which is ludicrous, *and* that self-ascriptions are based upon introspective observation. But when someone reports that he or she is afraid the basis for that report is not an introspective awareness of an internal state. If it was, then it should be feasible for someone to verify that their internal state was indeed one of fear. The suggestion that someone in the grip of fear might conduct an investigation of the contents of his mind in order to ascertain whether it was really fear that he felt rather than awe, or anger, is also ludicrous.

Similarly, since states of fear are individuated according to what is feared—so that a fear of *spiders* is distinguished from a fear of *snakes*—the presence of a certain internal state cannot be sufficient for the individuation of a particular fear. If it was then it would be conceptually possible for someone to observe that she is in a state of fearing a snake when there are no snakes to be seen, and even to confirm, at the same time, that she does not believe there are any snakes present. Such a case should be conceivable if we commonly accepted that fears are self-ascribed on the basis of introspection. However, I would suggest that in such a case we would doubt whether the person really understood what she was saying.

That said, we should note that Eliminativists do not need to suppose that first person ascriptions of psychological states are based upon introspective observation. As we saw in chapter 2, Paul and Patricia Churchland met the objection that first person uses of psychological concepts are not theoretical (because they are non-inferential), with the reply that although they may be non-inferential the 'semantic identity', to use the former's terms, of the concepts is 'fixed by a theory'.

However, if first person uses of psychological terms are not based on observational identification of states, events, or processes, of the kind that might figure in a theoretical taxonomy, then it must be the third person uses that are based on such observations. But, if my comments on third person ascriptions are correct, we cannot say that they individuate states of the sort recognized by even a pre-scientific taxonomy. As I have argued, the criteria for third person ascriptions are not criteria for individuating such states; a fact is reflected in what it does and does not make sense to ask and expect of a person. In the case of states of fear, anger, joy, and the like, what we observe are, firstly, states of the person rather than his brain and, secondly, states individuated only when the context of a person's reactions is acknowledged.

The conclusion I would urge is that insofar as terms like 'belief' or 'fear' can be used to refer, they are not used in the way that terms like 'liquid', 'phlogiston', or 'neuron' are used to refer. If this is the case then to say that the terms of the former variety are used to individuate entities identified within a theoretical taxonomy is to ignore the way that these terms are ordinarily used. Since folk psychology is supposed to be the framework within which these non-theoretical terms receive their ordinary usage, the implication is that folk psychology is not a theory.

Another way of phrasing this conclusion would be to say that the Eliminativists wrongly assume that predicates of the form 'believes that *p*' or 'remembers that *p*' are predicates picking out natural kinds. That is, they assume that the extensions of these predicates are populated by states, events, or processes sharing a common feature. In phrasing the conclusion in this way we can immediately see how it undermines the second assumption, required by Eliminativism to support the premiss that folk psychology constitutes a theory, *viz.*, that the folk are tacitly committed to a body of generalizations. The second assumption must presuppose that psychological predicates are used to pick out natural kinds if the predicates are to occur in nomic generalizations projecting them onto singular relations, between beliefs (and the like) and other psychological phenomena, and between beliefs and the environment and behaviour, for it is these singular relations that are to instantiate nomic relations posited by the folk generalizations. In other words, if beliefs do not have extensional properties (properties characterizing the extension of the predicate 'believes that *p*'), then it cannot be said that a particular ascription of a belief picks out something falling within the extension of the predicate ('believes that *p*') occurring in a nomic statement. But if, as I have argued, psychological predicates either are not used to refer to states (as in the case of 'belief that *p*'), or are not used to refer to states of a substance (as in the case of 'fear' or 'anger'), then they cannot be said to be kind predicates. And if they are not *kind* predicates then they are not projectible and, therefore, not suitable for picking out the subject matter of nomic statements constituent of a theoretical framework.

It should not go unnoticed that the foregoing creates more problems for Realist cognitive science. If there is to be a Realist science of cognition then the nomic statements sought by cognitive scientists will be couched in the ordinary psychological vocabulary. But if that vocabulary does not contain kind predicates then the search is pointless. However, there are further reasons to reject the assumption that there are psychological laws, as well as an aspect of the debate about this assumption, to which I should like to give brief consideration.

3 THE BODY OF GENERALIZATIONS

In 'Eliminative Materialism and the Propositional Attitudes' Paul Churchland writes;

> 'Each of us understands others, as well as we do, because we share a tacit command of an integrated body of lore concerning the lawlike relations holding among external circumstances, internal states, and overt behavior. Given its nature and functions this body of lore may quite aptly be called "folk psychology".' (P.M. Churchland, 1981, p.207)

Churchland gives, as an example of a law of folk psychology:

1) (x) (p) (q) [((x believes that p) & (x believes that (if p then q))) $\supset$ (barring confusion, distraction, etc., x believes that q)]

He suggests that such a law can be treated as on a par with laws from the physical sciences such as the classical gas law:

2) (x) (P) (V) (μ) [((x has a pressure P) & (x has a volume V) & (x has a quantity μ)) $\supset$ (barring very high pressure or density, x has a temperature of $PV/\mu R$)]

The supposed parity between 1) and 2) is taken by Churchland to provide grounds for rejecting the argument that folk psychology could not be replaced by a descriptive neuroscience because the former, but not the latter, has a normative characterization. The argument, which he attributes to Dennett and Popper, is that generalizations ranging over propositional attitudes constitute an idealization of rationality offering a standard to which the folk will approximate in their reasoning (ibid., pp.212&213). According to Churchland, the normativity of such generalizations is thought to be indicated by the fact that 'the regularities ascribed by the intentional core of FP are predicated on certain logical relations among propositions' but, as he sees it, the parallel between 1) and 2) shows that the fact 'is not by itself grounds for claiming anything essentially normative about FP' (Ibid., p.217). After all,

> 'the fact that the regularities ascribed by the classical gas law are predicated on arithmetical relations between numbers does not imply anything essentially normative about the classical gas law. And logical relations between propositions are as much an objective matter of abstract fact as are arithmetical relations between numbers.' (Ibid.)

Both the argument Churchland rejects, and his grounds for rejecting it, raise contentious issues. The issue I will concentrate on is that of the role 1) is supposed to play in the common-sense psychological framework. Both the argument, as Churchland renders it, and the rejection, suppose that there are indeed psychological laws governing the relations between environment, internal states, and external behaviour, only the argument seems to be that as *laws* of reasoning these must be distinguished from *laws* of nature, while the rejection is grounded on the claim that there is no relevant difference between the two. This claim, in turn, rests on the rendering of these, putative, laws in a way that rides roughshod over the ontological sensibilities of many philosophers, for it involves quantification over propositions.

However, I shall not exploit the contentiousness of this claim beyond observing that, in quantifying over both subjects and propositions, 1) casts beliefs as relations. This means that the folk, who have 'tacit command' over such laws, are unwittingly committed to seeing themselves and others as being somehow related to abstract objects. The implausibility of attributing such a commitment to the folk provides grounds for questioning the claim that statements like 1) play any role in our explanation of action, but even if we allowed that they did play such a role we should not ignore the fact that in 1) there is a conspicuous failure to quantify over beliefs themselves; a failure which undermines Churchland's insistence that the laws of folk psychology range over, amongst other things, internal states.

That said, the argument that 1) cannot be a law of nature because it has a normative character is misguided because 1) does not provide a norm of reasoning, or inference. This becomes clear as soon as we see that the entailment from the antecedent to the consequent is not a logical entailment. Churchland takes '...believes that *p*' to be a predicate forming expression from which we derive a determinate predicate by placing a singular term for a proposition into the argument position (ibid, pp.208&209 and see note 3, pp.222&223). By providing the predicate with a subject of predication we form a sentence constructed from two singular terms (namely '*x*', or, by universal instantiation, a constant which can replace it, and '*p*') and a predicate forming operator. Different sentences will result from substituting different singular terms for propositions. Thus '*a* believes that *p*' will differ from '*a* believes that (if *p* then *q*)' and, to make the important point, there will be no relation between the two sentences other than that they have the same predicate forming operator, the same subject, and the same letter (but not the same proposition) in the substitution place of the predicative expression, for although '*p*' occurs in both places, in the first it is a singular term whilst in the second it is a constituent of a singular term. There is, therefore, no logical relation between these two sentences nor between them and the sentence '*a* believes that *q*'. Thus we could render 1) schematically as ; $(P\&Q) \supset R$; where P,Q, and R are sentential variables. From this we can see that the relation between the antecedent and the consequent is not a relation of logical entailment, for the truth of $P\&Q$ does not logically entail the truth of *R*. Of course, the truth of *P* and $(P \supset Q)$ does entail the truth of *Q* but that is not the claim made by 1).

If 1) is a lawlike statement then the implication it asserts is not logical but nomological. The failure to distinguish between the two is, as we have seen, a prerequisite for the cognitive scientific enterprise and the claim that psychological laws have a normative character is a symptom of the failure, as I shall now try to show.

If 1) is, indeed, a lawlike statement then it is a statement that can be falsified by the existence of anomalous relations between beliefs—this should be the case if 1) is indeed a theoretical statement asserting nomic relations between items recognized by the folk psychological taxonomy, as the Eliminativists maintain. The inclusion, in 1), of the *ceteris paribus* clause 'barring confusion, distraction, etc.', should not preclude the possibility of falsification, for if it does then 1) will not be a law but a triviality of the form '*b* follows *a* except in those circumstances when it does not'. Now, the mistake in claiming that 1) has a normative character can be brought out by assuming that it does provide a standard of correctness and proceeding as follows: If *a* does not believe that *q*, even though he believes that *p* and that $p \supset q$, then this is can be due to one of two factors; either, the *ceteris paribus* clause *has not* been satisfied (that is, *a* has become confused, distracted, etc.), or the clause *has* been satisfied, and *a*'s reasoning is anomalous.

In the case of the first factor obtaining the law *has not* been falsified even though *x has not* made the correct inference and, in the second case, the law *has* been falsified, in which case it *does not* provide a standard of correctness for inference. That is, even if, in the first case, the law truly described nomological relations between beliefs, it could not distinguish between correct and incorrect inferential relations among them—if *a* believes *r* instead of *q* then, since 1) is true, this must be explained by the fact that the *ceteris paribus* condition has not been met rather than by anomaly among belief relations. So, if 1) is true it is not a standard of correctness. Since we began by assuming that the law was a standard of correctness, the falsification allowed in the second case straightforwardly falsifies the claim that 1) is a standard of correctness. The only other option left for those who want to say that 1) is normative is to claim that whatever fails to falsify it will count as compliance with a norm of reasoning. But to do this will be to allow that, provided one is confused, distracted, and so on, one can *correctly* infer just about anything one wants to. Apart from the fact that, if 1) turned out to be true, this would mean that there are no such things as incorrect inference relations among beliefs, it would require that the folk are entirely mistaken about what the take to be valid reasoning.

Thus, the argument that we cannot eliminate folk psychological laws and replace them with laws couched in the language of neuroscience because the former, but not the latter, are normative, is misguided. Laws, *qua* nomological generalizations, are not normative in the sense in which standards, or rules, of reasoning are normative because laws do not provide standards of correctness.

The question remaining is whether any role is left for statements like 1) in folk psychology. We may put the question thus: Is it the case that we appeal to such statements in explaining and predicting actions? To construct a particular case, if we know both that Gavin believes that if he leaves the keys in the ignition of his unlocked car overnight, then there is a good chance the car will be stolen, and that he knows he has done so, do we then explain his lack of surprise that the car has been stolen, when he goes to where he left it the next morning, by appeal to a law of the form of 1)? Well, if in appealing to 1) we are supposed to be appealing to the existence of a regularity between belief states of a certain kind, then we must ordinarily talk of beliefs, like those of Gavin, as states of some sort.[3] In the last section I argued that this is not how we ordinarily talk about beliefs. If we understood Gavin's belief, that he has left the keys in the ignition, to be a state of a substance then it would make sense for us to ask whether he had had his belief continuously or intermittently over night, and it would also make sense to ask him whether the belief state exists that morning. Indeed, we should have to ask him this in order to find out whether 1) was relevant in this circumstance, for if his belief states (relating to his car and the risk of having it stolen) had subsided then 1) could not explain his behaviour. If Gavin was surprised to find his car stolen the next morning we might suppose he has forgotten that he left the keys in the car, but we would not speculate as to the moment at which his state of belief ceased to be.

The point is that our ordinary way of talking about beliefs indicates that the criteria for their ascription are not criteria for ascribing states of a substance. It is no use claiming that, for example, we just assume that Gavin's states of belief have persisted throughout the night, because if this is to be a theoretical assumption then it would have to be the case that we accept some form of evidential basis for it. The fact that people often do assent to the same beliefs from day to day does not support the hypothesis that there are continuing *states* of belief, not only because the fact also supports the hypothesis that beliefs are created anew each day, but also because there can be no such hypothesis. This is because there is no way to establish the falsity of such a hypothesis—for how could we show that we do not continue to believe what we do when asleep or when not expressing our beliefs?—and where there are no criteria for falsity there are no criteria for truth either, and a statement

[3] We can lay aside the fact that 1) quantifies over propositions and believers, and not beliefs, and assume, for the sake of argument, a reformulation of 1) in which the existence of states of belief is implied. This removes the parity between 1) and 2), but that parity was debatable anyway. In quantifying over pressure, volume, and quantity, 2) requires the existence of what Churchland calls 'numerical attitudes', such as *n*kg/cm^2, *n*dm^3, or *n*mol, but not the existence of numbers (ibid., p.208). By contrast, 1) quantifies over propositions but not propositional attitudes.

which can be neither true nor false is not a hypothesis.[4] The *explanation* of Gavin's lack of surprise, then, is not provided by the positing of nomic relations between his belief states.

It is also important to note that 1) is unfit to play a role in *predicting* action. Predictions, like explanations, can be correct or incorrect, but 1) does not admit of such evaluation. We can begin by asking whether, on assuming that a believes that p, and that a believes that $(p \supset q)$, and that a is not confused, distracted, and so on, we could correctly predict that a will believe that q. Well, if we find out that a does believe that q then it will look to us as though we have deduced a correct prediction from 1) *until* we consider that, as a theoretical statement, 1) should be capable of producing incorrect predictions. But if we suppose that a does not believe that q we appear to have two alternatives. Firstly, we can say that our prediction was incorrect and that, provided it is an instantiation of the general statement, the general statement has been falsified, or, secondly, we can say that we were incorrect in assuming that a believed that $p \supset q$. I would suggest that the second alternative reflects what actually happens because it simply makes little sense to maintain that someone who *understands* that $p \supset q$ and believes that p may not believe that q.[5] Any apparent justification for maintaining that someone may understand the conditional and believe the antecedent but not the consequent is removed when we move from the general to the particular.

For example, if, one evening, Gavin tells us that he has left his keys in the car and that this means it will probably be stolen, then we might predict that he will not be surprised if it is. However, if the car is stolen and Gavin is surprised, the natural conclusion to draw would be either that he had forgotten he had left the keys in the car or, despite what he said, he did not really believe that the car might be stolen, or that he did not really understand that

[4] Note that the argument that beliefs are brain states, and that we *do* have criteria by which to judge whether brain states persist, simply begs the question; for we should not assume beliefs are brain states unless we already have criteria for the continued existence of beliefs as states.

[5] I am assuming, as Churchland does, that the implication involved is elementary.

leaving his keys in the car was sufficient for the likelihood of its being stolen. Thus, what 1) takes to be a causal consequent (namely (*x* believes that *q*)) is actually a criterion for correctly ascribing one, or both, of the antecedent conditions; (*x* believes that *p*) and (*x* believes that (if *p* then *q*)). It is this criterial feature of the relation between psychological ascriptions, behaviour, and the settings of that behaviour, that casts doubt on the claim that we abstract gross generalizations, like 1), from observations of the relations between belief states and behaviour, and then use these in explaining and predicting the occurrence of beliefs and behaviour. A generalization like 1) cannot be falsified by predictive failure because there is no clear division between what is predicted and the criteria for correctly individuating that upon which the prediction is supposedly based, *viz*. the beliefs given as the antecedent conditions of the generalization. There would be such a division only if, in the case of 1), we could determine that *a* believes that *p*, or that $p \supset q$, independently of what *a* goes on to do and say—that is, independently of whether *a* displays the belief that *q*—but unless we are able to identify belief states as states of a substance of some kind such a determination will be unavailable.

We might treat the formulation of 1) and generalizations like it, as grammatical in intent. If we do so we must be clear that 1) describes not a nomological relation but a grammatical one, for we might read it as saying that it is correct to say of someone that he believes both that *p and* that $p \supset q$ only if he believes that *q*. As such the generalization is a guide to what it makes sense to *say* about someone rather that a means of predicting or explaining his or her state of mind. This would mark the return to 1) of a normative status but *not* a nomological one, for the codification of the correct use of psychological ascriptions is not the description of nomologically necessary relations amongst states, events, or processes. Thus, if general propositions like 1) are not seen as theoretical statements, and are stripped of questionable quantifiers, then they might have a role as guides to, or arbiters in disputes about, the use of psychological expressions. However, such a role will be, at best, peripheral.

In summary, the conclusion I have argued for is that in attributing beliefs we do not tacitly employ a 'body of lore concerning the lawlike relations holding among external circumstances, internal states, and overt behavior'. Insofar as 1) typifies the generalizations to be found in the body of lore (see ibid., p.209 for other examples), this body of lore is certainly not implicated in the attribution of beliefs, fears, thoughts, and the like to ourselves and others. The normal commerce in psychological expressions does not presuppose that they are expressions denoting states, events, processes, or any other *genuinely* enduring phenomena to be found within subjects of ascription. Furthermore, the putative laws of folk psychology seem to be decidedly unsuitable for predicting behaviour since there is no room for predictive failure, from which it follows that predictive success is hollowed out to the point at which we should deny that the law had any predictive role. Of course, when a statement has no predictive role it does not deserve the epithet 'theoretical statement' and cannot, therefore, be attributed with an *explanatory* role. Indeed, the only way the 'body of lore' can be seen as explanatory is as a body of observations on the accepted use of psychological expressions. In this guise, Churchland's statements become explanatory to the extent that they describe some legitimate uses of those expressions and even proscribe some illegitimate ones—these being of the kind that would, if the statements *were* lawlike, falsify those statements—but they are not explanatory in the sense required of theoretical statements.

4 THE CONCLUSION

Section b) pressed for the rejection of the assumption that the psychological vocabulary (at least that part of it relevant to discussions of cognition) is referential in the way that a theoretical vocabulary should be. The last section argued that there is no body of lawlike generalizations from which we explain and predict cognitive behaviour. In effect, the two main premisses, from which Eliminativists conclude that folk psychology constitutes a theoretical framework, are ungrounded. Without the conclusion, however, an argument urging us to see the psychological vocabulary as borne of a false theory cannot be legitimately formulated.

There remains a consideration which we touched upon in the previous section but did not exploit. Churchland argued that generalizations of folk psychology are on a par with those of the physical sciences in order to show that, insofar as the former are characteristically normative, they are no more so that the latter. However, we found that demonstrating this parity was of little relevance to the debate because even if the generalizations of folk psychology were lawlike, they would not be normative. That is, the 'theoretical' generalizations of folk psychology would have no role to play in evaluating beliefs and behaviour as rational or irrational, for they could not act as norms and standards of rationality according to which beliefs and behaviour are judged as correct or incorrect. And yet, we do have such norms and standards for we do evaluate our words and deeds as correct and incorrect. So if psychological generalizations, and the vocabulary in which they are couched, are eliminated and replaced by those of a mature neuroscience, is it the case that we would retain norms and standards of reasoning and, for that matter, language use? In the following, albeit brief, answer to this question I think that what emerges is just how ill-conceived Eliminativism is.

One point that came to the fore in the discussion of normativity in chapter 6 was that a full account of normativity cannot make do just with notions of compliance or non-compliance with norms, for those notions by themselves do not create room for normative evaluations. To be *correct* in one's reasoning, for example, one must intend, or try, to reason correctly. To make a mistake or blunder in one's reasoning requires that one did not intend to reason badly, and a slip of the tongue when speaking is only a slip if one is trying to say something. Thus, what makes what one does correct, mistaken or a slip is not just its compliance or non-compliance with rules of reasoning or norms of expression, but the intention or desire to comply. Put simply, notions of correctness and incorrectness, error, and mistake, and their cognates, cannot be separated from notions of intention and purpose, and their cognates.[6]

Of course, an intention to speak, or to act rationally, is not to be understood as a mental act causing the speaking or acting; rather, as Wittgenstein puts it, 'an intention is embedded in its situation, in human customs and institutions' (1953, §337). To speak, or to act rationally, one must have learned to take part in activities commonly called 'reasoning' or 'speaking', and that one intends to reason and speak is shown by one's taking part. My claim is that it is the *concepts* of 'intention', 'purpose', 'desire' and so on, rather than states of intending, purposing, or desiring, that are bound to the concepts of 'correctness', 'mistake', and 'error'. But if the claim is right then the elimination of the psychological vocabulary would render obsolete the latter concepts and, therefore, normative evaluation itself.

It should be noted that the Eliminativist is not rescued from the incoherence of this consequence by the suggestion that replacing the psychological vocabulary with a neuroscientific one has the advantage of equipping us with a more precise means to presenting the explananda of folk psychology. The thought might be that, if psychological terms are required for an account of the normativity inherent in the explananda, then a more precise replacement vocabulary will give a more precise account of that normativity. However, two points need to be made in relation to this thought.

[6] Mechanisms do not *intend* to comply with rules and it is natural enough to speak of them functioning *correctly*, but only when they function as *intended* or *designed*. Thus, the normative evaluation of mechanistic functioning presupposes agency.

Firstly, replacing the psychological vocabulary with a neuroscientific one is incompatible with an account of normative phenomena, like language use and reasoning, unless the words of the new vocabulary are used in the same way as those of the old. If, for example, a friend told me, as we sat down to lunch, 'I am ravished' I might wonder whether he really means to say 'ravenous' or 'famished'; that is, I may wonder whether he has chosen the wrong word by confusing two which would be appropriate to the context. Ordinarily I would ask something like 'Don't you mean/intend to say "ravenous", or "famished"?' but perhaps, if the Eliminativists have their way, I will have to replace the intentional expression with one recognized by the neuroscientific theory. What is noticeable is that, whatever the replacement expression might be, it will need to function in the same way as the old one if it is do the same duty in raising the question of an error in locution. This being so we may well wonder just what has been gained by the replacement, for changes in orthography or phonetics achieve nothing if the new word is used in the same way. Insisting, for example, that we should always replace 'intention', or 'intent', with 'hippocampal state E' marks a change in the form of expression but not in concepts.[7]

Presumably our new vocabulary will offer us predicates of the kind that can be projected by the laws, governing behaviour, promised by neuroscience; laws that the old, anomalous folk psychology could not give us. This would seem to be the gain in the proposed replacement. However, leaving aside the objections to viewing the predicate matter of psychological sentences as things (like states, events and processes) to be subsumed by physical laws, it was norms and standards we were after rather than laws. The second point, then, is that what a science offers, and what it is often its purpose to offer, is a body of

[7] We might add that exchanging one substantive for another seems simple enough, but both 'intent' and 'intention' share the same Latin origin 'intendere' from which verbal, adverbial, and adjectival forms are derived. It will not suffice, therefore, to simply replace the substantival form of psychological terms, but it is doubtful that terms from the lexicon of neuroscience could have the other forms. What form could be given to the acceptable replacement of the expressions 'I intend to do so' or 'I did so intentionally', for example? Without such forms the notion of normativity we are considering could not be attached to action and speech.

nomological statements. But as we have seen, nomological statements do not provide norms to which normative evaluations attach. Lawlike regularities do not establish norms according to which actions can be deemed correct, appropriate, ill-judged, prudent and so on. A nomological statement can tell us what *ought* to happen in certain conditions, but when things turn out as the law predicts it is not because the relevant phenomena have successfully grasped what the law prescribes. When hypotheses deduced from the statement are predictive failures it is the statement that is incorrect rather than the events which do not conform to it.

The conclusion that we should eliminate the psychological vocabulary from explanations of behaviour, if this is to amount to the elimination of psychological *concepts*, would require that we can explain behaviour, whether linguistic or non-linguistic, without mention of norms affording normative evaluations. So, if the statement of the Eliminativists' conclusion is to make sense then we can do without evaluations of *nonsense* and *sense*. Of course, if one protects a statement from being judged to be nonsense by depriving it of the possibility of making sense, one must concede that one is not stating anything.

The foregoing suggests that if the use of the psychological vocabulary has a normative aspect, then that vocabulary is not one suited to the phrasing of nomological generalizations required by the Realist cognitivist scientist. But neither is the vocabulary susceptible to replacement by a neuroscientific one, as the Eliminative requires. This alone indicates that the idea of a cognitive science is not a cogent one.

REFERENCES

When reference is made to a reprint of a book or article below, unless stated otherwise, the page references in the text of the thesis are to the reprint rather than the original.

Anscombe, G.E.M. 1965. 'The Intentionality of Sensations: A Grammatical Feature'. In R. Butler (ed), *Analytical Philosophy*, second series, Oxford. Reprinted in *The Collected Papers of G.E.M. Anscombe, Volume Two: Metaphysics and the Philosophy of Mind*, Oxford: Basil Blackwell, 1981.

Armstrong, D.M. and Malcolm, N. 1984. *Consciousness and Causality: A Debate on the Nature of Mind*, Oxford: Basil Blackwell.

Baker, G.P. and Hacker, P.M.S. 1984. *Language, Sense and Nonsense*, Oxford: Basil Blackwell.

———1985. *Wittgenstein: Rules, Grammar, and Necessity*, Oxford: Basil Blackwell.

Beaney, M. (trans and ed), 1997. *The Frege Reader*, Oxford: Basil Blackwell.

Bechtel, W. 1987. 'Connectionism and the Philosophy of Mind'. In the *Southern Journal of Philosophy Supplement*, vol. 26, pp.17-41. Reprinted in Lycan (ed), 1990.

Block, N. 1986. 'Advertisement for a Semantics for Psychology'. In *Midwest Studies in Philosophy, X.* Reprinted in Stich and Warfield (eds), 1994.

Boghossian, P.A. 1989. 'The Rule-Following Considerations'. In *Mind*, vol. 98, no. 392, Oct. 1989.

———1991. 'Naturalizing Content'. In Loewer and Rey (eds), 1991.

Brentano, F. 1874. *Psychologie vom empirischen Standpunkt.* Translated as *Psychology from an Empirical Standpoint* by A.C. Ramurello, D.B. Terrell, and L.L. McAlister. Edited by L.L. McAlister, New York: Humanities Press, 1973.

Button, G., Coulter, J., Lee, J.R.E. and Sharrock, W. 1995. *Computers, Minds and*

Conduct, Cambridge: Polity Press.

Chomsky, N. 1957. *Syntactic Structures*. The Hague: Mouton & Co.

Christensen, S.M. and Turner, D.R. (eds), 1993. *Folk Psychology and the Philosophy of Mind*, Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Churchland. P.M. 1979. *Scientific Realism and the Plasticity of Mind*, Cambridge: Cambridge University Press.

———1981. 'Eliminative Materialism and the Propositional Attitudes'. In *The Journal of Philosophy*, vol. 78, pp.67-90. Reprinted in Lycan (ed), 1990.

———1988. *Matter and Consciousness*, Revised Edition, Cambridge, Mass.: MIT Press.

———1995. *The Engine of Reason, the Seat of the Soul: A Philosophical Journey into the Brain*, Cambridge, Mass.: MIT Press.

Churchland, P.S. 1986. *Neurophilosophy: Toward a Unified Science of the Mind-Brain*, Cambridge, Mass.: MIT Press.

Churchland, P.S. and Sejnowski, T.J. 1989. 'Neural Representation and Neural Computation'. In Nadel, Cooper, Culicover, and Harnish (eds), *Neural Connections and Mental Computations*, Cambridge, Mass.: MIT Press. Reprinted in Lycan (ed), 1990.

Clark, A. 1990. 'Connectionist Minds'. In *Proceedings of the Aristotelian Society*, vol. 90, pp.83-103. Reprinted in Macdonald and Macdonald (eds), 1995.

———1993. *Associative Engines: Connectionism, Concepts, and Representational Change*, Cambridge, Mass.: MIT Press/Bradford Books.

Conant. J. 1998. 'Wittgenstein on Meaning and Use'. In *Philosophical Investigations*, vol. 21, no. 3, July 1998.

Davidson, D. 1963. 'Actions, Reasons, and Causes'. In the *Journal of Philosophy*, vol. 60. Reprinted in Davidson, 1980.

———1967. 'Causal Relations'. In the *Journal of Philosophy*, vol. 64, pp.691-703. Reprinted in Davidson, 1980.

———1970. 'Mental Events'. In L. Foster and J.W. Swanson (eds), *Experience and Theory*, Amherst, Mass.: University of Massachusetts Press.

———1980. *Essays on Actions and Events*, Oxford: Oxford University Press.

———1984. *Inquiries into Truth and Interpretation*, Oxford: Oxford University Press.

———1993. 'Thinking Causes'. In J. Heil and A. Meil (eds), *Mental Causation*, Oxford: Clarendon Press.

Dennett, D.C. 1969. *Content and Consciousness*, London: Routledge & Kegan Paul.

———1971. 'Intentional Systems'. In the *Journal of Philosophy*, vol. 68, pp.87-106.

———1975. 'Brain Writing and Mind Reading'. In Gunderson (ed), 1975. Reprinted in Dennett, 1978.

———1978. *Brainstorms: Philosophical Essays on Mind and Psychology*, Brighton, Sussex: Harvester Press.

———1978a. 'Artificial Intelligence as Philosophy and as Psychology'. In M. Ringle (ed), *Philosophical Perspectives on Artificial Intelligence*, New York: Humanities Press. Reprinted in Dennett, 1978.

———1981. 'True Believers: The Intentional Strategy and Why It Works'. In A.F. Heath (ed), *Scientific Explanation*, Oxford: Oxford University Press. Reprinted in Dennett, 1987.

———1983. 'Styles of Mental Representation'. In the *Proceedings of the Aristotelian Society*, vol. 83. Reprinted in Dennett, 1987.

———1984. 'Cognitive Wheels: The Frame Problem of AI'. In C. Hookway (ed), *Minds, Machines, and Evolution*, Cambridge: Cambridge University Press.

———1987. *The Intentional Stance*, Cambridge, Mass.: MIT Press/Bradford Books.

———1987a. 'Reflections: Instrumentalism Reconsidered'. In Dennett, 1987.

———1987b. 'Reflections: The Language of Thought Reconsidered'. In Dennett, 1987.

Dretske, F.I. 1981. *Knowledge and the Flow of Information*, Oxford: Basil Blackwell.

———1986. 'Misrepresentation'. In R. Bogdan (ed), *Belief: Form, Content and Function*, Oxford: Oxford University Press. Reprinted in Sitch and Warfield (eds), 1994.

———1991. 'Dretske's Replies'. In B.P. McLaughlin (ed), *Dretske and His Critics*, Cambridge, Mass.: Basil Blackwell, 1991.

Dummett, M. 1973. *Frege: Philosophy of Language*, London: Duckworth and Co, Ltd.

Eccles, Sir J. 1987. 'The Effect of Silent Thinking on the Cerebral Cortex'. In B. Gulyás (ed), *The Brain-Mind Problem: Philosophical and Neurophysiological Approaches*, Louvain.

Field, H.H. 1978. 'Mental Representation'. In *Erkenntnis*, vol. 13, no. 1, pp.9-61. Reprinted in Stich and Warfield (eds), 1994.

Fodor, J.A. 1975. *The Language of Thought*, Cambridge, Mass.: Harvard University Press.

———1980. 'Methodological Solipsism Considered as a Research Strategy in Cognitive Psychology'. In *The Behavioural and Brain Sciences*, vol. 3, no. 1. Reprinted in Fodor, 1981.

———1981. *Representations: Philosophical Essays on the Foundations of Cognitive Science*, Brighton, Sussex: Harvester Press.

———1983. *Modularity of Mind: An Essay on Faculty Psychology*, Cambridge, Mass.: MIT Press/Bradford Books.

———1984. 'Semantics, Wisconsin Style'. In *Synthese*, vol. 59, pp.231-250. Reprinted in Fodor, 1990.

———1985. 'Fodor's Guide to Mental Representation: The Intelligent Auntie's Vade-Mecum'. In *Mind*, vol. 94, Spring 1985, pp.55-97. Reprinted in Fodor, 1990.

———1987. *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*, Cambridge, Mass.: MIT Press/Bradford Books.

———1987a. 'Making Mind Matter More'. In *Philosophical Topics*, vol. 67, no. 1, pp.59-79. Reprinted in Fodor, 1990.

———1990. *A Theory of Content and Other Essays*, Cambridge, Mass.: MIT Press/Bradford Books.

———1990a. 'Psychosemantics or: Where Do Truth Conditions Come From?'. In Lycan (ed), 1990.

———1998. *Concepts: Where Cognitive Science Went Wrong*, Oxford: Oxford University Press.

Fodor, J.A. and Lepore, E. 1991. 'Why Meaning (Probably) Isn't Conceptual Role'. In *Mind and Language*, vol. 6, no. 4. Reprinted in Stich and Warfield (eds), 1994.

Fodor, J.A. and McLaughlin, B.P. 1990. 'Connectionism and the Problem of Systematicity: Why Smolensky's Solution Doesn't Work'. In *Cognition*, vol. 35. Reprinted in Macdonald and Macdonald (eds), 1995.

Fodor, J.A. and Pylyshyn, Z.W. 1988. 'Connectionism and Cognitive Architecture: A Critical Analysis'. In *Cognition*, vol. 28. Reprinted in Macdonald and Macdonald (eds), 1995.

Frege, G. 1879. *Begriffsschrift, eine der arithmetischen nachgebildete Formelsprache des reinen Denken*, Halle: L. Nebert.

———1884. *Die Grundlagen der Arithmetik: eine logisch-mathematische Untersuchung über den Begriff der Zahl*, Breslau: W. Koebner. Selections reprinted and translated in Beaney (trans and ed), 1997. Page references are to the original.

———1892. 'Über Begriff und Gegenstand'. In *Vierteljahrsschrift für wissenschaftliche Philosophie*, vol. 16, pp.192-205. Reprinted and translated in Beaney (trans and ed),

1997. Page references are to the original.

Goldstein, L. and Slater, H. 1998. 'Wittgenstein, Semantics and Connectionism'. In *Philosophical Investigations*, vol. 21, no. 4.

Gunderson, K. (ed), 1975. *Language, Mind, and Knowledge: Minnesota Studies in the Philosophy of Science VII*, Minneapolis: University of Minnesota Press.

Hacker, P.M.S. 1990. *Wittgenstein: Meaning and Mind, Part 1: Essays*, Oxford: Basil Blackwell.

Hume, D. 1740. *A Treatise of Human Nature: Book One*, D. Macnabb (ed), London: Wm Collins Sons & Co., 1962.

Ince, D.C. (ed), 1992. *Collected Works of A.M. Turing: Mechanical Intelligence*, Amsterdam: North-Holland.

Kitcher, P. 1996. 'From Neurophilosophy to Neurocomputation'. In R.N. McCauley (ed), *The Churchlands and their Critics*, Oxford: Basil Blackwell.

Lahav, R. 1989. 'Against Compositionality: The Case of Adjectives'. In *Philosophical Studies*, vol. 57, pp.261-279.

Lewis, D. 1972. 'Psychophysical and Theoretical Identifications'. In the *Australasian Journal of Philosophy*, vol. 50, pp.427-446.

Locke, J. 1690. *An Essay Concerning Human Understanding*, abridged and edited by R. Wilburn, London: J.M. Dent and Sons, 1947.

Loewer, B. and Rey, G. (eds), 1991. *Meaning in Mind: Fodor and his Critics*, Oxford: Basil Blackwell.

Lycan, W. G. (ed), 1990. *Mind and Cognition: A Reader*, Oxford: Basil Blackwell.

Lyons, W. (ed), 1995. *Modern Philosophy of Mind*, London: J.M. Dent & Sons Ltd.

Macdonald, C. 1995. 'Introduction: Connectionism and Eliminativism'. In Macdonald and Macdonald (eds), 1995.

Macdonald, C. and Macdonald, G (eds). 1995. *Connectionism: Debates on Psychological Explanation, Volume 2*, Oxford: Basil Blackwell.

Malcolm, N. 1958. 'Knowledge of Other Minds'. In the *Journal of Philosophy*, vol. 55, no. 23, pp.969-978.

Marr, D. 1982. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*, San Francisco, Cal.: W.H. Freeman.

Mill, J.S. 1889. *An Examination of Sir William Hamilton's Philosophy*, sixth edition, London.

Millikan, R.G. 1989. 'Biosemantics'. In the *Journal of Philosophy*, vol. 86, no. 6. Reprinted in Stich and Warfield (eds), 1994.

———1991. 'Speaking up for Darwin'. In Loewer and Rey (eds), 1991.

Mills, S. 1993. 'Wittgenstein and Connectionism: a Significant Complementarity?'. In C. Hookway and D. Peterson (eds), *Philosophy and Cognitive Science: Royal Institute of Philosophy Supplement: 34*, Cambridge: Press Syndicate of the University of Cambridge.

Minsky, M. 1981. 'K-Lines: A Theory of Memory'. In D. Norman (ed), *Perspectives on Cognitive Science*, Norwood, N.J.: Ablex

———1981a. 'A Framework for Representing Knowledge'. In J. Haugeland (ed), *Mind Design*, Cambridge, Mass.: MIT Press/Bradford Books.

Palmer, A. 1988. *Concept and Object: The Unity of the Proposition in Logic and Psychology*, London: Routledge.

———forthcoming. 'Contingent Identities and Category Differences'. In *Revue Internationale de Philosophie.*

Piaget, J. 1971. *Structuralism*, London: Rouledge & Kegan Paul.

Place, U.T. 1956. 'Is Consciousness a Brain Process?'. In the *British Journal of Psychology*, vol. 47. Reprinted in Lyons (ed), 1995.

Proudfoot, D. 1997. 'On Wittgenstein on Cognitive Science'. In *Philosophy*, vol. 72, pp.189-217.

Putnam, H. 1960. 'Minds and Machines'. In S. Hook (ed), *Dimensions of Mind*, New York: New York University Press. Reprinted in Putnam, 1975a.

———1961. 'Some Issues in the Theory of Grammar'. In *Proceedings of Symposia in Applied Mathematics*, vol. 12, pp.25-42. Published by the American Mathematical Society. Reprinted in Putnam, 1975a.

———1973. 'Philosophy and Our Mental Life'. Presented at a symposium on 'Computers and the Mind' at the University of California (Berkeley). Published in Putnam, 1975a.

———1975. 'The Meaning of "Meaning"'. In Gunderson (ed), 1975.

———1975a. *Mind, Language and Reality: Philosophical Papers, Volume 2*, Cambridge: Cambridge University Press.

Quine, W.V.O. 1951. 'Two Dogmas of Empiricism'. In the *Philosophical Review*, Jan. 1951, vol. 60.

———1956. 'Quantifiers and Propositional Attitudes'. In the *Journal of Philosophy*, vol. 53, pp.177-187.

Ramsey, W., Stich. S.P., and Garon, J. 1990. 'Connectionism, Eliminativism, and the Future of Folk Psychology'. In *Philosophical Perspectives*, vol. 4. Reprinted in Macdonald and Macdonald (eds), 1995.

Robinson, G. 1992. 'Language and the Society of Others'. In *Philosophy*, vol. 67. no. 261, pp.329-341.

Rorty, R. 1965. 'Mind-Body Identity, Privacy, and Categories'. In the *Review of Metaphysics*, vol. 19, pp.24-54. Reprinted in Christensen and Turner (eds), 1993.

Russell, B. 1940. *An Inquiry into Meaning and Truth*, London: Allen and Unwin Ltd.

Ryle, G. 1949. *The Concept of Mind*, Middlesex, England: Penguin University Books.

Sayre, K. 1987. 'Cognitive Science and the Problem of Semantic Content'. In *Synthese*, vol. 70, pp.247-269. Reprinted in D. Cole, J. Fetzer, and T. Rankin (eds), *Philosophy, Mind, and Cognitive Inquiry: Resources for Understanding Mental Processes*, Dordrecht, The Netherlands: Kluwer Academic Publishers, 1990.

Searle, J.R. 1980. 'Minds, Brains and Programs'. In the *Behavioural and Brain Sciences*, vol. 3.

———1983. *Intentionality: An Essay in the Philosophy of Mind*, Cambridge: Cambridge University Press.

Sellars, W. 1965. 'The Identity Approach to the Mind-Body Problem'. In the *Review of Metaphysics*, vol. 18, pp.430-451.

Smart. J.J.C. 1959. 'Sensations and Brain Processes'. In the *Philosophical Review*, vol. 68, pp.141-156. Reprinted in Lyons (ed), 1995.

Smolensky, P. 1988. 'On the Proper Treatment of Connectionism'. In *Behavioural and Brain Sciences*, vol. 11. Reprinted in Macdonald and Macdonald (eds), 1995.

———1991. 'Connectionism, Constituency and the Language of Thought'. In Loewer and Rey (eds), 1991. Reprinted in Macdonald and Macdonald, 1995.

———1995. 'On the Projectable Predicates of Connectionist Psychology: A Case for Belief'. In Macdonald and Macdonald (eds), 1995.

———1995a. 'Reply: Constituent Structure and Explanation in an Integrated Connectionist/Symbolic Cognitive Architecture'. In Macdonald and Macdonald (eds), 1995.

Stich, S.P. 1983. *From Folk Psychology to Cognitive Science: The Case Against Belief*, Cambridge, Mass.: MIT Press.

Stich, S.P., and Warfield, E.A.1995. 'Reply to Clark and Smolensky: Do Connectionist

Minds Have Beliefs?'. In Macdonald and Macdonald (eds), 1995.

Stich, S.P., and Warfield, E.A. (eds), 1994. *Mental Representation: A Reader*, Oxford, UK. and Cambridge, USA.: Basil Blackwell.

Thornton, T. 1998. *Wittgenstein on Language and Thought: The Philosophy of Content*, Edinburgh: Edinburgh University Press.

Turing, A.M. 1948. 'Intelligent Machinery'. National Physical Laboratory Report. Reprinted in Ince (ed), 1992.

——— 1950. 'Computing Machinery and Intelligence'. In *Mind*, Oct. 1950, vol. 59, no. 236, pp.433-460. Reprinted in D.C. Ince (ed), 1992.

Wilkes, K. 1981. 'Functionalism, Psychology and the Philosophy of Mind'. In *Philosophical Topics*, vol. 12, no. 1, pp.147-167.

———1991. 'The Relationship Between Scientific Psychology and Common Sense Psychology'. In *Synthese*, vol. 89, pp.15-39. Reprinted in Christensen and Turner (eds), 1993.

Winch, P. 1960. *The Idea of a Social Science and its Relation to Philosophy*, second edition, London: Routledge and Kegan Paul.

Wittgenstein, L. 1922. *Tractatus Logico-Philosophicus*, translated by C.K. Ogden, London: Rouledge & Kegan Paul Ltd.

———1953. *Philosophical Investigations*, translated by G.E.M. Anscombe, Oxford: Basil Blackwell.

———1969. *On Certainty*, edited by G.E.M. Anscombe and G.H. von Wright, translated by Anscombe and D. Paul, Oxford: Basil Blackwell.

———1976. 'Cause and Effect: Intuitive Awareness'. In *Philosophia*, vol. 6, nos. 3-4, pp.391-445, translated by P. Winch. Reprinted in J. Klagge and A. Nordmann (eds), *Philosophical Occasions, 1912-1951*, Indianapolis and Cambridge: Hackett Publishing

Company.

———1981. *Zettel,* second edition, edited by by G.E.M. Anscombe and G.H. von Wright, translated by Anscombe, Oxford: Basil Blackwell.