USE OF THESES

This copy is supplied for purposes of private study and research only. Passages from the thesis may not be copied or closely paraphrased without the written consent of the author.
Emotion & Evolution

Paul Edmund Griffiths,
Dept. of Philosophy,
Research School of Social Sciences,
Australian National University.

A thesis submitted for the degree of Doctor of Philosophy of The
Australian National University.

June 1988
This thesis is my own work, written while I was a Research Scholar in the Research School of Social Sciences, Australian National University. To the best of my knowledge, all sources have been acknowledged.

P.E Griffiths,
10th June 1988.
"He who admits on general grounds that the structure and habits of all animals have been gradually evolved, will look at the whole subject of Expression in a new and interesting light."

Contents

Acknowledgments IX

Abstract X

Chapter one: Introduction

1.1. Emotion & the Philosophy of Mind 1

1.2. The Role of Folk Psychology in our Account 3

1.3. The Role of Propositional Attitude Psychology 6

Chapter Two: The Cognitive Theory

2.1. Origins of Cognitivism 8

2.2. Straightforward Cognitivism & its Problems 11

2.3. Adding Desires to the Analysis 15

2.4. Beyond Beliefs and Desires 19

2.5. What Would A Cognitive Taxonomy of Emotions Look Like? 24
Chapter Three: The Psychoevolutionary Theory

3.1. Darwin's Work on Emotional Expression 29
3.2. Modern Work on Emotional Expression 33
3.3. Darwin's Theory of Emotion Evolution 37
3.4. The Modern Psychoevolutionary Theory — The Case for Innate Elements in the Emotional Response 44
3.5. The Modern Psychoevolutionary Theory — The Case for Affect Programs 52
3.6. The Neural Basis of the Affect-Programs 60
3.7. Limitations of the Affect-Program Approach 65

Chapter Four: The Constructionist Theory

4.1. What is Constructionism? 72
4.2. The Problem of Sincerity 85
4.3. The Problem of Inappropriate/Sinful Emotions 96
4.4. Problems of Universality, Innateness & Passivity 101
4.5. The Real Role of Constructionism 103
4.6. Summary of Conclusions

Chapter Five: Overview: A Multi-Vocal Theory of Emotion

5.1. The Need for a Multi-Vocal Theory

5.2. The Multi-Vocal Theory

5.3. The Basis of the Folk Taxonomy

5.4. The Paradigm Status of the Affect-Programs

5.5. Alternative Taxonomies

Chapter Six: The Attribution of Emotions

6.1. Historical Introduction

6.2. Attributing Cognitive States

6.3. Attributing Affect-Program States

6.4. Attributing Constructed States

6.5. Summary

Chapter Seven. Moods and Character Traits

7.1. Introduction

7.2. Lormand's Theory
7.3. Moods as Higher-Order Dispositions 159
7.4. Moods as Higher-Order Functional States 161
7.5. Support from the Folk Conception 165
7.6. Support from the Neurosciences 166
7.7. Conclusions 171

Appendix I. Experimental Details. (Chapter 3)


I.II. Ekman, Sorensen & Malmstrom (1970) — Component Analysis. 175

Appendix II. The Sociobiological Option. (Chapter 5)

II.I. Sterelny’s Argument 178

II.II. Reply to Sterelney 180

II.III. A Counterexample to Sterelney 182

Appendix III. A Theory of Sensations. (Chapter 6).

III.I. Introduction 183

III.II. Neuropsychology & Sensations 187
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>III.III. The Conservative Inference</td>
<td>190</td>
</tr>
<tr>
<td>III.IV. The Radical Inference</td>
<td>191</td>
</tr>
<tr>
<td>III.V. Epiphenomenalism</td>
<td>194</td>
</tr>
<tr>
<td>III.VI. Neuropsychological Qualia</td>
<td>198</td>
</tr>
<tr>
<td>III.VII. Some Recalcitrant Intuitions</td>
<td>200</td>
</tr>
<tr>
<td>III.VIII. Conclusions</td>
<td>205</td>
</tr>
<tr>
<td>Bibliography</td>
<td>207</td>
</tr>
</tbody>
</table>
Acknowledgements

I am deeply indebted to all my supervisors for their consistently constructive criticisms. At various times these supervisors have included Professor J.J.C Smart, Professor P Forrest, Dr P Pettit and Prof F.C Jackson. I am particularly grateful to Dr K Sterelny, who was my primary supervisor until his departure from the School, and who has continued to provide superogatory comments on my work. As well as being prepared to re-read unreasonably large numbers of drafts, Dr Sterelny helped me to see that a naturalised approach to emotions, in line with recent work in other areas of philosophical psychology, offers the only chance for substantial philosophical progress in this area.

As well as formal supervision, I have benefited greatly from discussion with other members of the School. I might mention, amongst others, Dr Jerry Gauss, Mr David Braddon-Mitchell, and Mr John Fitzpatrick. The School's continual flow of distinguished visitors has also been of great assistance, and I am grateful for comments on portions of my work by several visitors. Names which spring immediately to mind are Prof. Isaac Levi and Prof. Michael Smith.

Earlier versions of chapter three, chapter seven and appendix III have been read at Australasian Association for Philosophy conferences in 1986, 1987 and 1988 respectively. A version of chapter three was read to a seminar of the Department of Traditional and Modern Philosophy at the University of Sydney. Various parts of the thesis have been read to "work in progress" seminars in the School, and to meetings of the Australian National University Philosophy Society. In all cases, my ideas have been clarified by the ensuing discussion.

Finally, I should thank my fiancée Ms Marie-Louise McBride, for putting up with three years of composition.
Abstract

In my introductory chapter I briefly outline the present state of philosophical emotion theory. I state my position on the status and future of "folk psychology", and the future of traditional propositional attitude psychology.

The next three chapters of the thesis expound and criticise some current theories of emotion. In chapter two I outline the prevalent "cognitivist" orthodoxy. I present a series of problems for this theory, and show how successive refinements to the theory have attempted to meet them. In the last section of the chapter I discuss the relationship between a cognitive taxonomy of emotions and the folk taxonomy.

In chapter three I give an account of the philosophically neglected psychoevolutionary theory of emotions. I describe Darwin's remarkable work in this field, and compare it to the present state of play. In sections four and five of the chapter I outline an "affect-program" model of emotions, and connect it with current work in other areas of the philosophy of mind.

In chapter four I discuss the constructionist theories of emotion which have been taken up in recent years by certain philosophers. I criticise many of the current accounts of constructionism, and produce my own, substantially revised version, which I claim has been purged of many muddles, conflations, and controversial commitments.

Chapter five defends the core contentions of the thesis. I argue that none of the three theories discussed are adequate accounts of the whole domain, and that each should only properly apply to certain subdomains of emotional phenomena. I call this a "multi-vocal" theory of emotion, in contrast to the traditional insistence on univocal theories. Univocal theories, I argue, are merely Procrustean.
In chapter six I provide a cognate account of emotion attribution, stressing the diversity of modes of attribution that corresponds to the diversity of "emotional" phenomena.

Chapter seven provides a theory of mood and trait phenomena. Although independently motivated, the theory turns out to dovetail neatly with my account of affect-program states.

The appendices deal with matters referred to in the thesis, but which would interrupt the flow of my argument if included in the main body of the text. Appendix I gives details of experiments referred to in chapter three. Appendix II defends the uses of sociobiology suggested in chapter five. Appendix three gives a theory of sensations referred to in chapter six and elsewhere.

Format of the Thesis

I have used three levels of headings. First, the thesis is divided into chapters, numbered with Arabic numerals 1—7. Next, each chapter is divided into subsections numbered decimally, 1.1, 1.2, 1.3, etc. Finally, simple underlined heads, such as that used to introduce this paragraph, are inserted at various points to aid the exposition. These are not listed in the table of contents, or used for cross-referencing.

Appendices to the thesis are numbered with Roman numerals, I—III. They are divided into decimally numbered sub-sections, I.I, I.II, etc, in a manner analogous to the chapters. Each appendix is linked to the particular chapter where the questions it deals with are most frequently referred to.
Figures and tables are also numbered decimally. Figures in chapters are labeled Figure 3.1, Figure 3.2, etc. Tables in Appendix I are labeled Table I.I, Table I.II, etc.
Chapter One: Introduction.

1.1. Emotion and the Philosophy of Mind

Emotions have been treated in a rather cavalier fashion in the mainstream philosophy of mind. For the last twenty-five years, most effort has been devoted to showing that emotional phenomena can be dealt with by a conventional propositional attitude psychology modelled on our folk explanations of intentional action. This mopping up operation goes under the name of the cognitive theory of emotion.

The cognitive theory is primarily concerned to combat the so-called "feeling theory" of emotions. The feeling theory holds that emotions are inner experiences, characterised by a quality and intensity of sensation. Their identity depends on this quality, and the inherent pleasantness or unpleasantness of that quality causes us to act in various ways. The cognitivist thinks this picture pernicious, as well as false. It is said to prevent us from subjecting emotions to rational criticism, and to conduce to the view that emotions are beyond our control.

The feeling theory either originates in, or was reinforced by, eighteenth century empiricist theories of mental contents. Locke, Berkeley and Hume standardised the view that the contents of the mind are sensations, some perceptive, and some interoceptive. The emotions are just one class among the many classes of sensation. But the feeling theory of emotions has proved more enduring than the general picture of the mind as a collection of sensations. The average non-philosopher is strongly attached to the idea that emotions are sensations. It seems merely common sense. I do not think the average person has any such attachment to
the view that thoughts in general consist of concatenations of sensations or images. Gilbert Ryle, in "The Concept of Mind" testifies that in his day the adherent of the feeling theory could count on: "the consent of most philosophers and psychologists to the view that emotions are internal, or private, experiences. ...turbulences in the stream of consciousness" (p81). It is a matter of history how effectively Ryle and other behaviourist thinkers were able to extirpate the general sensational theory of the mind, both in psychology and philosophy, but the feeling theory of emotion has proved far more resilient. It has been continuously advocated in the works of psychological writers, such as Silvan Tomkins (1962, 1963, 1979). In recent years, it has resurfaced in the philosophical literature, in the work of Michael Stocker (1983, 1987).

I shall argue in chapter six and appendix three that the supposedly definitive case made out against the feeling theory by Ryle, Anthony Kenny, and many other authors, runs up against some very obstinate psychological and neurological data. But whatever the final verdict, it is important to understand how obsession with this question has distorted the philosophical discussion on emotion. The attempt to get cognitivism up and running, and the derivative debates on the rationality and moral status of emotions, have totally dominated the literature. Alternative approaches, such as the psychoevolutionary theory, have been left in almost total obscurity. The publication in the last five years of articles on the constructionist theory has rather broadened the debate, but even here the old concern with the dichotomy of feeling and cognition has been apparent. The philosophical constructionist, rather than challenging the cognitivist orthodoxy, sees herself as an ally against the evils of the feeling theory! In consequence, the issues between the profoundly individualistic cognitive theory of emotions and a theory which stresses the place of emotion within the social system remain almost unexplored.

At least part of the reason for this impoverishment of the philosophical debate is the tradition of concern with mental state attribution, rather than with substantial
theories about the mental. In most of the standard philosophical literature, questions concerning the nature of emotions are inextricably intertwined with discussions of how they are attributed. The cognitivist's opposition to the feeling theory does not spring from any view about the substantive nature of emotions, but from general worries about the epistemology of the mental. The identity of emotion states cannot depend upon quality of sensation, because mental terms must be publicly definable to be part of our language. People cannot recognise their emotion on the basis of a sensation, because of the supposed impossibility of a private language. Whatever we may think of these arguments, it must be admitted that little knowledge about emotion itself is manifested in them.

Alternative theories of emotion, such as Darwin's, and those emerging from our developing understanding of the mind/brain, do not primarily address themselves to these epistemological questions. Their first aim is to give a substantive account of what emotions are. They look for an account of the mind/brain processes which underly emotional phenomena. The standard philosophical literature neglects this question. It is the substantive theory of emotion which will concern us in this thesis. I hope it will become apparent that the traditional philosopher's attempt to give a theory of emotion attribution without first giving a substantive account of emotions was seriously misguided. Empirical research reveals mechanisms that are left unconsidered in the traditional epistemological debate, and, furthermore, the emotions turn out to be a diverse collection of phenomena, with correspondingly diverse modes of attribution.

1.2. The Role of Folk Psychology in our Account.

Throughout the thesis I refer to our "folk" discourse about emotions. I assert that there is a folk scheme for taxonomising emotions, and I describe what I claim to be the uses of mood terms in folk discourse. The nature of this folk discourse is
a matter of philosophical controversy, and I should at least sketch my relationship to that debate. I take it to be generally admitted that there is a non-theoretical discourse about minds. By calling it non-theoretical I don't imply that it is radically different in kind from our theoretical discourse, merely that it is unregimented, and not subject to any consciously applied methodology. I admit that at the heart of this discourse there lies a model of action explanation which relies on normative principles of rationality. I deny, however, that this is all there is to that discourse. Folk discourse exists to facilitate the everyday interactions between people, and those people only roughly approximate to rational systems. A pragmatically useful understanding of human beings must involve ideas about their reflexes, failures of rationality, common modes of malfunction, and so on. Common sense discourse about the mental therefore embodies entities and principles over and above those required for rational action explanation. This fact will be brought out in our account of the folk understanding of emotions.

I argue in chapter five that the folk-taxonomy of emotions, and indeed, the category of emotion itself, is unlikely to be preserved in a scientific psychology. These categories fail to enter into the generalisations of any useful theory. This suggestion may be resisted by two groups of people. The first are those who believe that there is some difference in kind between the explanations offered in folk discourse, and those which may come from a scientific understanding of the mind/brain. As should have been suggested by my opening remarks about folk-psychology, I do not share this view. The issues between such a view and my own do not directly concern the emotions, and will not be discussed here. The second group are those who are impressed by the pragmatic success of folk discourse, and expect some of its structure to be preserved in a scientific psychology. I think that this view can be reconciled with mine, if I can persuade such people that our pre-scientific understanding of love, envy, and so on is not as good as our understanding of other aspects of the mental, such as the rationalisation of actions by beliefs and
Chapter One: Introduction

desires. This is a fairly plausible claim. A simple distinction will make it yet more plausible. I argue in the thesis that we understand people's emotions despite, rather than with the aid of, our emotion categories. Generalisations using those categories only hold up because they are supplemented by a case by case understanding of the exceptions. For example, it seems almost truistic to assert that you can only be afraid of dangerous things, and that if you really appreciate the danger of something, you ought to fear it. The relationship to danger seems almost constitutive of our understanding of what it is to be afraid. But to get that relationship to hold up we have to tack onto our understanding of fear such ad hoc additions as cowardice, bravery, and the acute fear of harmless things such as worms. I hope to show in the thesis that this is because "fear" denotes a number of very different phenomena. I shall argue that this is true of most of our emotion categories. This fact is reflected in our tendency, when discussing emotions, to insist that there are many "kinds" of whatever folk category is in question.

Suppose I can convince at least the second type of objector that the various folk categories of emotion need revision. What of the claim that the category of emotion itself should be revised? One consideration which should have some weight here is the very recent arrival of this category into our language. The word "emotion" is transferred from application to physical movements to "movements of the mind" in the late seventeenth century. It slowly takes over from "passion" as the general term for fear, love and so on in the eighteenth century, and its adjectival form "emotional" is a nineteenth century construction, as are most other terms based on the same root. A change of extension, as well as connotation, is involved in the move from "passions" to "emotions", and the merely "emotional" is a still wider (and vaguer) category. So the "emotional" is not a category deeply and irremovably embedded in our understanding of persons. Once we cease to allow the common name to captivate us, it is far from obvious that the suprise engendered by a loud

1 See, for example, the dates of the examples of these uses listed in the Oxford English Dictionary.
noise, the yuppy's envy of the successful old guard, and the love of Dante for Beatrice, form a single class of mental events. Their diversity is all too apparent. This diversity will appear still greater upon closer acquaintance. Closer examination will also show that univocalism, the insistence that each new theory be stretched to cover all emotional phenomena, has led to the overextension and distortion of what might otherwise have been valuable theoretical approaches to certain subdomains of emotional phenomena.

1.3. The Role of Propositional Attitude Psychology

In chapter two of the thesis I discuss the cognitivist view that emotions are sets of propositional attitudes, and conclude that some of them may be. This creates something of a difficulty. The thesis aims to perform the philosophical groundwork for the scientific investigation of the events in the mind/brain which underly emotional behaviour. Chapter two would therefore seem to commit me to the view that propositional attitudes are psychologically real. Such a view is, at the very least, controversial. Certain philosophers, notably Stephen Stitch and Paul and Patricia Churchland, have argued that propositional attitude psychology as we know it will have no place in a developing science of the mind/brain, and will fall into abeyance. Others have argued that propositional attitude psychology constitutes a humanly necessary feature of our thought. They would expect it to be preserved as a separate realm of discourse, unaffected by developments in scientific psychology. In either case, the portion of our account which essentially involves propositional attitudes would have no place in a scientific story about emotional phenomena. This portion of my account may encounter further problems because of the many outstanding puzzles about the ascription of content to mental states, and the precise nature of the explanatory relation between propositional attitude ascriptions and behaviour.
I believe we can legitimately sidestep these important issues for the purposes of the present enquiry. Whatever the future holds, the arguments offered for isolating certain emotional phenomena from the scope of other accounts we shall offer will still apply. There will still be a class of events which I will here call cognitive emotions, and it will still be true that they cannot be accounted for either by the affect-program theory or the constructionist theory. The general structure of our account should therefore be preserved. Whatever theory comes to take the place of propositional attitude psychology will be faced with the task of explaining what underlies this class of phenomena. Since none of these accounts exist in any concrete form, we cannot say how they would set about this.

Our present discussion will show how a traditional philosophical psychology can use content bearing states akin to folk propositional attitudes to explain these phenomena. Even if such philosophical accounts turn out to be wildly off the scientific mark, our efforts will not have been useless. This is the sort of psychology which is presumed by most current accounts of emotion, and our examination of how such a psychology works will give us a better understanding of those theories and their shortcomings.
Chapter Two: The Cognitive Theory of Emotions.

2.1. Origins of Cognitivism

In this chapter I want to bring out the shortcomings of the cognitive theory of emotions. Cognitivism has been the dominant philosophical theory of emotion for the last twenty-five years. During this period it has been modified in various ways by various schools and individuals. The structure of the chapter will reflect this fact. In the first section I give a general characterisation of cognitivism, and refer to some of its origins in the work of post-Wittgensteinian philosophers such as Anthony Kenny. In section two I take the simple but influential cognitive theory espoused by R.C Solomon and use this as a foil to outline some basic problems with cognitivism. In section three I outline attempts to meet these problems by adding desires to the cognitive analysis, and in section four attempts to solve them by taking emotional contents to be something other than beliefs and desires. Finally, in section five, I consider the relationship between the cognitive taxonomy of emotions and the folk taxonomy, and find this more complex than most cognitivists have supposed. It will become clear that even if cognitivism had succeeded on its own terms, it would not have been able to give a full account of the emotions.

William Lyons characterizes cognitivism in the following manner:

"in general a cognitivist theory of emotion is one that makes some aspect of thought, usually a belief, central to the concept of emotion and, at least in some cognitive theories, essential to distinguishing different emotions from one another." (Lyons 1980 p33.)
But the first half of this characterization is a little misleading. All emotion theories give cognition a large role. Even a leading non-cognitivist psychologist like Robert Plutchik is happy to assert that:

"the existence of any emotion supposes the prior occurrence of a cognition or evaluation." (Plutchik 1980 p10.)

But he remains a non-cognitivist because he insists that, nevertheless,

"an evaluation is not an emotion, evaluations are a part of the total process". (Plutchik 1980 p9.)

Cognitivism makes evaluation not merely central to emotion but more or less constitutive of it. It does this by giving identity conditions for emotions in terms of the cognitions that they involve.

Cognitive psychologists refer to these cognitions as "evaluations", "appraisals" or "judgments." It is often unclear just what these terms refer to. Factual judgments, assessments of value to the subject, moral or quasi-moral judgments, and even expectations, are all sometimes cited as underlying emotions. Philosophers on the other hand have given a great deal of attention to the nature of the states with which they propose to analyse emotions. In paradigmatic philosophical cognitivism, an emotion type is distinguished by a particular type of propositional attitude, or attitudes, which it involves, and a particular occurrence of an emotion is distinguished by the occurrence of that attitude towards a particular intentional object. In other words, emotions are to be individuated by the content clauses of associated propositional attitude ascriptions. What sorts of propositional attitudes these are to be varies from theorist to theorist.
In the next section I shall describe the work of one particular theorist in whom paradigmatic cognitivism is relatively straightforwardly embodied. First, however, I shall consider a precursor of the standard view. The rise of cognitivism can be attributed in no small part to Anthony Kenny's 1963 work, "Action, Emotion and Will". Kenny makes no direct reference to the view that emotions should be analysed using sets of propositional attitudes, let alone identified with those sets. But it is not hard to perceive how this view could grow out of what he does explicitly say. Kenny is concerned to insist that emotions have formal objects. That is to say, they are intensional states directed onto objects of some particular class. The defining property of the class is what makes the emotion the emotion that it is. Thus, the formal object of envy is another's good, and that of gratitude a good done to oneself. The linkage to propositional attitudes should be obvious, as it is to Kenny himself. When he cites Aristotle as saying that anger is a desire to revenge what is believed to be an insult (p193), he accepts this as equivalent to his own formulations in terms of formal objects. We can state this equivalence as follows: when an emotion has the class of A's as its formal object, this is to say that the bearer of the emotion must believe or desire that A if he is to have that emotion. So Kenny's formulation in terms of formal objects yields fairly easily the claim that certain propositional attitudes are necessary for certain emotions. The corresponding sufficiency claim is somewhat harder to derive. Talk of formal objects lends itself naturally to formulating restrictions on the proper occasions of emotions, that is, to stating necessary conditions for emotions. It is less natural, however, to say that whenever there is a suitable formal object there must be an emotion. The obvious way to formulate this claim is to say that the formal object suited to an emotion exists whenever we have beliefs and desires which specify such a formal object. This makes it possible to state the
sufficiency claim as follows: whenever a person has propositional attitudes which specify the formal object of an emotion, then they have that emotion.

Kenny himself doesn't make this sufficiency claim, but there are two reasons why a cognitivist should make this claim, once he couches his or her claim in terms of propositional attitudes. The first reason might appeal most to Kenny. If you think that emotions are in some sense logically related to their objects, then having an appropriate emotion is the rational thing to do. If you believe that you are in danger, and desire not be in danger, it is appropriate to be afraid, and irrational not to be. If you fail to be afraid, this gives others some reason to doubt whether you really know what is happening, or really desire to be out of danger. Thus, in an ideal model, our emotions are always in agreement with the available formal objects, and real life cases where we fail to emote are simply failures of rationality, akin to our failure to obey deductive closure with regard to our beliefs. The second argument for the sufficiency claim is simply that as part of the general reductive project for propositional attitudes, emotions should be analysed out as beliefs and desires. To be angry just is to believe yourself wronged and to desire vengeance. Most cognitivists have been fairly strongly influenced by this general reductive project and there is no place in most of their accounts for any irreducible propositional attitudes other than belief and desire.

2.2 Straightforward Cognitivism & Its Problems

I shall now outline a particularly simple and strident cognitivist theory put forward by Robert Solomon in "The Passions" (1977) and elsewhere. Solomon straightforwardly identifies emotions with evaluative beliefs:

"my anger is (authors emphasis) that set of judgments .... an emotion is an evaluative (or a normative) judgment". (p185)
As a disciple of Sartre's, Solomon is concerned to show that his account is more complex than this and hints at a constructivist element:

"an emotion is a basic judgment about ourselves and our place in the world, the projection of the values and ideals, structures and mythologies...." (p185-6.)

But his efforts in this direction are not backed up in his actual theory. On a true constructivist view an emotion is primarily a pattern of behaviours learnt by members of a culture, and expressing a conception about what is "appropriate" behaviour in a given, socially defined situation. I discuss these views at length in chapter four. Solomon only flirts with this sort of view and it has no modifying effect on his identification of emotions with simple evaluative judgments:

"my embarrassment is my judgment to the effect that I am in an exceedingly awkward situation .... my sadness, my sorrow and my grief are judgments of various severity to the effect that I have suffered a loss'. (p186)

Problems for Straightforward Cognitivism

Solomon's account is an excellent starting point for a critique of cognitivism. It has been widely enough attended to be no mere straw man and yet is simple enough to be open to almost all the criticisms that more subtle cognitivist theories have tried to avoid. It gives us a chance to display the persistent, underlying problems with a cognitivist approach. These may be listed under six heads.
1. Objectless Emotions

States such as depression, elation, and anxiety are generally thought to be capable of "clinical" instances where they have no object and thus no content. The standard cognitivist reply to this objection is to deny that there are such states. Lyons, for example claims that generalized depression requires the judgment that things are pretty bad and the object of that state is things generally.

2. Reflex Emotions

The judgments which cognitivists take to be at the heart of emotions are often very different from ordinary judgments. The fear of earthworms commonly co-occurs with the conviction that earthworms are harmless. The "judgment" underlying the fear is one we would hotly deny making. Such "judgments" appear to be made more rapidly and less consciously than many judgments, and are to varying extents insulated from the rest of our beliefs. Solomon, following Sartre, argues that all such cases are examples of self deceit, but this is frankly implausible.

3. Unemotional Evaluations

Identifying emotions with evaluative judgments gives us far too many emotions. Many smokers believe smoking to be dangerous but smoke without fear. There are also many evaluative judgments that are never likely to be the contents of emotions. An apple sorter makes such judgments thousands of times every day.

4. Judgments Underdetermine Emotions

Lyons gives the example of the judgment that Ashkenazy is a fine pianist. This might give rise to envy or to admiration or perhaps to other emotions, or
perhaps to no emotion at all. The occurrence of these various emotions need not prima facie involve the judge holding any further beliefs about Ashkenazy. Such cases may lead us to believe that the evaluative beliefs associated with an emotion do not exhaust its content or that emotions are not individuated by their contents.

5. Physiological Responses

The cognitive theory neglects the physiological aspects of emotion. Emotional responses are characterised by at least four classes of physiological changes, facial expressions, skeletal/muscular changes such as flinching, expressive vocal changes and autonomic nervous system changes such as adrenalin release and heartrate changes. There is empirical evidence that at least some of these responses are the invariable concomitants of certain emotions\(^1\). It is also clear that our recognition of emotions is to a large extent dependent on our observation of these factors\(^2\), yet cognitivism appears to leave them entirely puzzling.

6. Emotions via Imagination

Michael Stocker(1987) has pressed the view that we can have full blown emotions by imagining suitable objects. Here the beliefs and desires of the cognitive analysis are ones we explicitly don't have! The obvious cognitivist comeback would be to insist that in such cases we don't actually have emotions, but only imagine we have, and that if we ever become genuinely emotional, we must have confused fantasy and reality. I think this defence can be sustained quite convincingly against most anecdotal evidence. Nevertheless,

\(^1\) See Chapter 3, Section 2.
\(^2\) See Chapter 6, Section 3.
Stocker's is surely the prima facie plausible interpretation of these cases, and the fact that it has to deny that interpretation is one more point against cognitivism.

2.3. Adding Desires to the Analysis

We might call the simple cognitive theory we have just criticised the belief theory. Many theorists have believed that the problems just outlined can be avoided by including desires in our analysis. J. Robinson (1983) argues that the contents of emotions behave more like desires than like beliefs. She lists the following desire like features of emotions. First, emotions, like desires, come in degrees of intensity, and these degrees are not obviously a matter of the strength of certain beliefs. The strength of my envy is not just the strength of my belief that someone possesses the envied good. Secondly, when I have mixed emotions with regard to an object this does not seem to be because I am not sure whether it is, say, dangerous or exciting. The object may clearly be both dangerous and exciting and my fluctuation between one emotion and another seems more akin to a fluctuation between inconsistent desires. Thirdly, emotions are highly tolerant of such inconsistencies in the same way that desires are. It is not clearly irrational to have conflicting emotions or desires, in the way that it is to have conflicting beliefs.

Joel Marks' Theory

Joel Marks (1982) offers a full belief/desire theory of emotion. The theory itself is very simple. Emotions are belief/desire complexes characterised by strong desire. How does this allow him to cope with our six problems? Marks has nothing to say about problem one, the objectless emotions. Like any cognitivist he must simply deny that there are any. Similarly, I think he would have to give
the usual cognitivist reply to objection six, emotions via imagination. But he can hope to do better than most cognitivists with problem two, the reflex emotions. Robinson has noted that it is characteristic of desires to tolerate contradictions. Marks hopes to use this and another feature of desires to explain how we can both fear spiders and wish that we did not fear them, since we know them to be harmless. The other feature Marks uses is that desires are resistant to change in the light of information contradicting that which originally caused us to have them. These two features allow Marks to analyse Tom's fear of spiders as the result of his strong desire to avoid spiders, and his belief that he has encountered one. He can explain the robustness of Tom's fear in the face of the information about the harmlessness of the spider, and his desire not to fear it by the two principles of the robustness of desires in the face of new and hostile information and the ability of desires to tolerate contradictions.

Marks hopes to solve the problem of the underdetermination of emotion by judgements and the problem of unemotional judgements (problems 3&4) simply by adding appropriate desires to the conditions for the possession of the various emotions. In answer to problem five, the problem of physiological response, he argues that we already know that strong desires are accompanied by physiological disturbance, and the disturbance which accompanies emotion is just an instance of this.

**Problems for Marks' View**

Marks' main worry about his own theory is that it is too liberal. He considers the case of a person who has a strong desire to go to Paris on August 18. Ordinarily this might not be called an "emotion" (although using the word "yearning" might change people's minds). But it is plausible that someone with
such a desire might experience a rise in heart rate when considering their trip or become excessively concerned about it, and thwarting their desire would lead to sadness or depression. Marks thinks that these features are sufficient to make the belief desire set qualify as an emotion. I do not think that can give any great weight to our intuitions about such cases. Our use of the word "emotion" is very loose and the boundaries of what constitutes an emotion are not at all clear. Whether we accept Marks's defence will probably depend upon our inclination to accept his theory on other grounds.

Marks should also give some thought to defending himself against the accusation of excessive chauvanism. The positive emotions resulting from the satisfaction of desire are a good example. At least some instances of happiness are cases where a person has no strong unfulfilled desires left, since they have just been satisfied. One might, I suppose, cook up some connected desires for these cases, such as the desire that what has been attained be not lost, but it seems back to front to suggest that the strength of our satisfaction depends on the strength of our wish not to have that satisfaction taken away. Furthermore, there are plenty of circumstances where what has been achieved can never be reversed, and these seem as likely to give rise to positive emotions as any others.

The major problem with Marks' account, however, is his notion of desire. When we speak of desires as they occur throughout philosophical psychology, we are thinking of the states which interact with beliefs to produce rational action, and which are attributed on the basis of their efficacy in explaining behaviour. Most often we think of such states as functional or dispositional. But this notion will not serve Marks's purpose. Marks wants to use the desire component to account for the physiological and phenomenological effects of emotions. He argues that we know that strong desires are accompanied by
physiological effects and distinctive phenomenology, and should therefore expect emotions to be accompanied by such effects if they are composed in part of strong desires. To get this argument to go through, Marks needs to use "desire" in the sense in which it denotes urgent, consciously manifest states. Such states are phenomenologically salient, and are labelled "yearnings", "longings", or "burning desires". Despite the problems of the famous James/Lange theory it seems reasonable to assume that this phenomenology is connected with the perception of the physiological effects, such as adrenalin release, raised heartrate, etc, which accompany these states. When we use "desire" in this sense the strength of a desire is taken to mean its degree of phenomenological salience - just how "burning" or "painful" it is. But in the more usual philosophical use of "desire" the strength of a desire is taken to be its motivational effect, as revealed in actual or counterfactual behaviour. It is a commonplace that these two kinds of strength do not correlate at all well. Our sexual desires are generally far more salient phenomenologically than our moral desires, yet most of us manage to behave reasonably well when the two conflict.

The disassociation between these two senses of "strong desire" allows us to offer Marks a fairly unpalatable dilemma. Suppose Marks opts to use desire in its belief/desire explanatory sense. He will have a concept which seems suitably independent of the emotions, and suitably central to at least some philosophies of mind, to provide an attractive notion with which to reduce the seemingly puzzling phenomena of emotion. But his account will do little to explain the physiological and phenomenological effects of emotions. Suppose, on the other hand, that he uses the phenomenological concept of desire. His explanatory project will be more likely to succeed, but what of his reductive project? The notion used to reduce the emotions seems as problematic as the emotions themselves. We have no clear taxonomy of these phenomenological desires,
and in understanding their relationships to qualia, and to physiological changes, we encounter the same problems found in relating these things to emotions. In fact, these "desires" seem to be a variety of affect, along with moods, emotions and the rest. Marks account will (almost) be one which reduces emotions to beliefs and emotions!

2.4. Beyond Beliefs and Desires.

In his section I want to discuss the theory of William Lyons in his book "Emotion"(1980), and that of Michael Stocker in his article "Emotional Thoughts"(1987). Both present telling arguments against cognitivism, and offer to go beyond beliefs and desires in analysing emotions. I shall argue that Lyons' account collapses into the belief/desire theory, and that Stocker's account, while a considerable advance, fails by giving a mystical account of its central new addition to the analysis.

**Lyons' Theory**

William Lyons holds a "causal-evaluative" theory of emotions. Emotions are evaluations of an object which cause physiological disturbances. The "causal" element is added simply to deal with the problem of unemotional evaluations mentioned above. It is "evaluations" which are central to Lyons' theory. Emotions are taxonomised by the type of evaluative attitude they involve. To fear x is to evaluate x as dangerous, to love it is to evaluate it as loveable (p62-3). I have said that Lyon's account collapses into the belief/desire theory, but this is not quite correct. There are two strands to Lyon's account of evaluation. Only one of these experiences the collapse, but I shall argue that the other is simply a matter of mysterious handwaving. The first is represented at pages 58-60. Here Lyons says:
"I tell myself that the dog is liable to take a large piece out of me....I would prefer that it did not, in a very strong sense of prefer, that is, if I evaluate the dog's approach as threateningly dangerous,"

and again:

"to evaluate is not to gain knowledge, but to relate something...to some rating scale."

It seems quite reasonable to accept Joel Marks' interpretation of these remarks (Marks 1982). To evaluate x as dangerous is to believe it capable of harming you and to desire that it not do so. Further support for this interpretation can be gleaned from Lyons' assertion in chapter six that the "formal object" of an emotion is the category into which it must be evaluated to produce that emotion. The formal object of fear is the dangerous-and-disliked, because that is how it must be evaluated to be feared. I argued in section 2.1 that talk of formal objects is equivalent to talk of propositional attitudes towards objects, and so by transitivity Lyons' talk of evaluations must be equivalent. To evaluate something as dangerous and disliked is to have a state whose formal object is the dangerous and disliked, which is, as argued above, to believe something dangerous and desire it to be otherwise. The first strand of Lyons' account, then, collapses into the belief desire view.

The other strand in Lyons' account of evaluation is his talk of "seeing as". Suppose a climber sees a rock face and judges it to be dangerous. Lyons claims that he would then go on to evaluate the rock face by "seeing it as" either exciting or frightening. This process decides whether we feel excitement or fear. It has to be said that this sounds suspiciously like the substitution of metaphor for actual theory. The phrase "seeing as" and associated phrases occur at several points but there is no sustained attempt to explicate them. If
some more detailed account were to be culled from Lyons of this new, sui generis mental process, we would no longer be able to collapse his account into that of the belief/desire cognitivist. But there are few materials in Lyons' work to base such an account on.

Lyons' theory, however, is not merely the evaluative theory, it is the causal-evaluative theory. Lyons does not call an evaluation emotional unless it causes some sort of physiological disturbance. Lyons is presumably sensitive to the same pressures as Marks, but instead of arguing that we can expect physiological effects as a concomitant of his account, he explicitly requires their presence. I doubt whether it is true that all sound English emotion words have physiological disturbance as part of their assertion conditions. We recognise all sorts of quite calm states as emotional. Nevertheless, Lyons probably has picked up an important feature of our usage. We are happier to consider, say, "hope" as an emotion if it seems to cause substantial physiological disturbance. I hope that my friend will have a pleasant weekend, but it is stretching things a bit to say that I have an emotion here. Someone might reply that this is because I don't really hope here, I'm just being polite, but this is not convincing. In the scenario described I am quite sincere, and would probably be prepared to expend some resources to procure my friend a pleasant weekend. That seems enough for hope, but perhaps not quite enough for an emotion. I hope the source of these intuitions will become clearer in the next chapter, and in chapter five.

Despite this concession concerning the importance of physiological changes, Lyons is insistent that emotions are attributed and taxonomised purely on the basis of the evaluations that accompany, or compose, them. When I see someone getting upset, says Lyons, I have to infer their mental state by the usual behavioural/explanatory means, and discover how they are evaluating the
situation in order to discover the significance of their upset. But this view of attribution and taxonomy turns out to be straightforwardly, empirically false. In chapter three I discuss a large body of results on the cross-cultural recognition of emotions from facial displays. The fact is that the information we get from this source is often enough to override our best guesses arrived at by other means. As for the evaluation based taxonomy, we shall see in the last section of this chapter that cognitive differentiation alone doesn't give us the folk-taxonomy we know and love. The way we do get that taxonomy has a lot to do with the physiological effects of emotion. Lyons' attempt to go beyond cognitivism by using the physiological effects of emotion comes to very little, because he makes so little use of his new resource. In all essential respects Lyons remains a cognitivist.

**Stocker's Theory**

I now turn to Stocker's 1987 account. Stocker presents a powerful case against the conventional cognitivist, relying largely on objections four and six of our earlier list. (Emotions are underdetermined by my beliefs, and I can have real emotions as the result of imagining things.) Stocker wants to show that while emotions do involve entertaining propositions, the attitude we have to those propositions need not be one of belief, or indeed of any evidential or truth-aiming kind. Take the following example of underdetermination. When Stocker is in an aircraft he claims he can oscillate between fear and confidence, without changing his beliefs about air safety. Stocker takes this to show that having the emotion is a matter of having some attitude other than belief to propositions about the aircraft's crashing. Stocker tries to make the same point with cases of of imagining oneself into an emotion. Suppose Stocker is in a dark room. He knows that snakes are not often found in suburban Melbourne, but imagining that one is coiling away silently in that dark corner, he finds himself becoming
afraid. Stocker takes this to show that the way in which a proposition activates one's emotions need not be by becoming more probable or better grounded. Later in the article he runs a parallel argument for desires. Stocker sums up his position quite well in the following way:

"the claim that emotions have cognitive content is not usefully put by holding they involve beliefs. Often, they involve only emotionally held thoughts. Similarly, emotions need not involve values and desires which the agent really has. They, too, can be held only emotionally."

(p59)

Clearly, the big question is what it is to "emotionally hold" a belief or desire. Stocker is not at all clear about this. At various points he talks of "attending to", "entering into", "living" and "taking seriously" these propositional contents, but little illumination is to be had from this proliferation of tired metaphors. Stocker himself admits:

"I find the notions of entering into and living a thought very difficult to characterise. But, as I trust the cases presented here have shown, they are not so difficult to illustrate." (p64)

The burden of this last remark is that Stocker takes his examples to be a conclusive refutation of cognitivism, and supposes that a new propositional attitude is the only conceivable solution. He is not too worried about his inability to immediately characterise that attitude, since he is sure it must exist, and that his examples capture it. But, as I shall show in the next chapter, the introduction of notions such as modularity and informational encapsulation, will allow us to handle these examples without positing this mysterious new attitude, and in a way that meshes in well with the general programme of the modern philosophy of mind. Stocker has correctly appreciated the
shortcomings of cognitivism, and has appreciated that the problems can be solved only by introducing into the analysis something unique to emotions, and quite separate from our usual stock of action rationalising beliefs and desires. Unfortunately, he chooses to go for some notion of a qualitatively different kind of mental content, a notion which from its very introduction is irremediably mysterious. In the next chapter, we shall see how to elucidate the idea of specifically emotional mental contents without using any such problematic notion as an emotional attitude to propositions.

2.5 What Would a Cognitive Taxonomy of Emotions Look Like?

Supposing that all the above criticisms were answered, and that a cognitive theory were credible, the theory would yield a series of analyses of emotion types, using the content based taxonomy presumed to exist for propositional attitudes. Emotion types, on the cognitive view, are to be imagined as content schemas. But the cognitive analyses of emotion found in the literature are content schemas of a very special sort. This is because these analyses try to explicate the categories of folk psychology. In giving a cognitive analysis of the folk category "fear", we mustn't be so fine grained as to differentiate the state of thinking x is a man eater and wanting to run from thinking x is armed and wanting to cycle away. This need for rough grainedness becomes even more apparent when we consider different environments and cultures, whose inhabitants may face an utterly different range of stimuli with an utterly different set of background beliefs, and still be afraid. We need a "schema" which is fitted by many, radically different propositional attitudes.

To see how such a schema works, we need to distinguish this sense of "schema for propositional attitudes" from various more familiar senses. In one very
common sense, a propositional attitude schema strictly determines what is believed or desired of something, but leaves open what it is that is believed or desired about in this way. Suppose we take "P believes a is red", considered as a strict schema. If this schema applies to John and a Lady Margaret blazer, then John believes of that blazer precisely that it is red (and not, for example, that it is whatever colour such blazers usually are). Clearly, this isn't the sort of "schema" we have in mind when we offer "P believes a is dangerous and wants to avoid a" as a schema for fear. We want "is dangerous" to be substitutable in particular contexts by "is a man-eater", "will tell the press". But allowing the substitutability of co-referring expressions won't solve the problem. The problem here is not the traditional point about opaque and transparent belief contexts. The change from "Believes a will tell the press", to "Believes a is a man-eater", is not a change of co-referring expressions within one belief context. These are wholly different beliefs. They are, in fact, two different ways of believing that a is dangerous. In the cognitive analysis of fear, "believes a is dangerous" is a vague description covering a large range of possible propositional attitudes.

This fact about cognitivism has perhaps been overlooked because there are two ways to read a sentence like "P believes that a is dangerous." First, it could be read as a strict schema for a vague belief. On this reading, P believes that there is something dangerous about a, without specifying the respect in which a is dangerous. You don't have to have this vague belief every time you are afraid. It should, on the proposed analysis, follow from whatever you do actually believe, but there is no reason to insist that you must draw this conclusion in order to be afraid. There are, of course, senses of "believes" so weak that they guarantee that you believe the deductive closure of your beliefs, but these are
not the senses relevant to the cognitive analysis of emotions. On any more substantial reading of "belief", there is no reason to suppose that whenever a person believes a lion is about to eat them, they must infer that the lion is dangerous before they can act.

The second way to read "P believes a is dangerous" is as a vague schema, which doesn't tell us precisely what P believes about a, but merely tells us that whatever he believes amounts, among other things, to believing a to be dangerous. It is left open whether P believes that a is a vicious political opponent, or that a is the spirit of his ancestor inhabiting a tribal mask. Clearly, it is the second sense that we want if we want to give a specification of all the various combinations of propositional attitudes that could, in some situation, constitute fear.

There is nothing special about fear in this regard. Consider these purported analyses of other emotions. Hope is the state of believing a situation to be possible and strongly desiring that situation. Envy is to be analysed with at least the following conditions. A envies B iff (1) A believes that there is a state of affairs S which benefits B. (2) A desires that such a state of affairs not exist. (3) A does not consider a world in which -S obtains superior to one in which S obtains except on the supposition that B benefits from S. Clearly, there are many very different sets of propositional attitude which could fulfill the requirements of these proposed analyses. A cognitive analysis is a rough specifications of an aggregate of different sets of attitudes, each of which could plausibly be supposed to be sufficient for hope or envy.

This suggests something very interesting about the supposedly "cognitive" taxonomy of emotions offered by the cognitivists. If the folk emotion categories

---

1 The cognitivist cannot be using these senses, or he would be forced to claim that if it follows logically from what you believe that you are subject to a danger you wish to avoid, then you are afraid, a claim which many a flustered person can refute.
don't correspond to particular sets of propositional attitude types, how does the cognitivist explain the way in which various sets of propositional attitude types are grouped together under one folk-psychological rubric? With "fear", we may be able to give some plausible specification of "danger" that doesn't render itself circular by referring to fear. But envy? It is hard to see what leads the cognitivist to group together the whole diverse range of mental states that meet the three broad conditions just outlined, other than the fact they they all constitute envy!

This observation is extremely significant. It means that the folk-emotion categories do not emerge smoothly from a cognitive taxonomy of propositional attitudes. It is not generally true that there is some cognizable property like dangerousness that all the objects of a particular emotion have in common. The cognitive category which is supposed to analyse an emotion must be derived from a prior understanding of that emotion. True, all the members of this category have a class of logical consequences in common (that's how we identify the category), but so do a vast number of overlapping categories, all equally good from a cognitive point of view, but nearly all of which aren't emotions.

So even if cognitivism were perfectly successful, it would not explain very much about the emotions. It would remain utterly dependent on prior knowledge of the folk taxonomy in deriving its own taxonomy, and could not offer an explanation of the fact that we are commonly jealous and envious, rather than schmealous and schmenvious. In the next few chapters, and particularly chapter five, it will become clear that there are important stories to be told about why many emotions are as they are, stories that are intimately connected with substantial accounts of what these states actually are. But these stories are generally not cognitive in nature. The theories that provide these
explanations will also provide us with ways of taxonomising emotions which are of more use to our explanatory project than the cognitive one.
3.1 Darwin's Work on Emotional Expression

The psychoevolutionary theory has its roots in Darwin's 1872 book "The Expression of the Emotions in Man and Animals". Darwin presented an account of the evolutionary principles which he thought could explain the emotional responses of the various species. He also attempted a detailed account of the responses themselves, concentrating on those of man. Darwin believed that human expressions of emotion were, to a large extent, adaptations of directly functional responses which had existed in our forebears, and which had been retained for their expressive value after their functional significance had declined. For example, he saw the baring of the teeth in angry humans as descended from a more primitive response, in which it served as a preparation for attack:

"with mankind some expressions, such as the bristling of the hair under the influence of extreme terror, or the uncovering of the teeth under that of furious rage, can hardly be understood, except in the belief that man once existed in a much lower and animal-like condition. The community of certain expressions in distinct though allied species as in the movements of the same facial muscles during laughter by man and by various monkeys, is rendered somewhat more intelligible, if we believe in their descent from a common progenitor." (Darwin 1872 p12)

Darwin's methods of investigation anticipate many of the experiments conducted by modern theorists. At the beginning of his book he cites the sources of evidence from which he drew his views on the expression of emotion. He examined the expression of emotion first in children and secondly in lunatics, in both of whom he hoped to find uncorrupted, natural expressions of emotion. In giving this weight to
Chapter three: The Psychoevolurionary Theory

In his use of Duchenne's and others' photographs Darwin anticipated the two most common forms of research on emotion used by later theorists. We may call these the component approach and the judgment approach. In the component approach the theorist attempts to arrive at hypotheses concerning the role of various facial muscles in the expression of the emotions. Darwin used Duchenne's hypotheses and other evidence, to form his own components hypotheses. He then tried to test these with judgement tests. The invention of the judgment test was perhaps Darwin's most original methodological contribution. In judgement tests, subjects are shown still photographs, films or drawings, of what are thought to be clear emotional expressions and are asked to state which emotion they think is expressed in the picture. Darwin used Duchenne's photographs for the first ever judgment test. He was able to show that certain of the photographs elicited very general agreement and took these to be reliable.
expressions of their associated emotions. Regrettably, Darwin only used the judgment test in England. Its main use in later theory has been in demonstrating the cross cultural recognition of emotional expression.

Darwin did attempt to obtain information about the expression of emotion by members of other cultures as his fourth source of evidence. He did this by means of questionnaire. He asked missionaries, traders and other residents in foreign cultures to answer questions on the facial components to express emotion by members of those other cultures. He obtained results from Australia, the Pacific, India, the two American continents and Africa. Unfortunately, Darwin's questionnaire suggested which facial components he expected to represent which emotions. He thus made it very likely that his observers would read the desired results into their observations.

As a potential fifth source, Darwin studied the expression of emotion in painting and sculpture but he found this of very little use. Later research has demonstrated that the use of drawings in judgment tests as opposed to photographs or films produces abnormal and uninteresting results (See Ekman 1973 p197-200)

Finally Darwin studied emotional expressions in animals. He hoped to find evidence of the evolutionary continuity of expression among several species. In this aim he was at least partly successful, though it would be inappropriate to give a survey and assessment of his specific results here.

Darwin himself did not conceive of his work as explaining the evolution of the emotions, or as giving an account of what emotions are. His aims were relatively limited when compared to modern psychoevolutionary theories. He claimed to explain only why emotions are expressed in the way that they are. This is because Darwin held a "feeling theory" of emotions. He accepted that emotions themselves are sensations, as when he endorsed Spencer in the following passage:
"Mr Herbert Spencer has drawn a clear distinction between emotions and sensations, the latter being "generated within our corporeal frame". He classes as feelings both emotions and sensations." (Darwin 1872 P:27)

Modern emotion theorists, however, tend to regard emotions as syndromes, meaning that they have many elements, and that no elements is obviously sufficient to constitute an emotion on its own. It may well be inappropriate to regard a sensation of fear which has no behavioural or physiological consequences as an instance of the emotion of fear. This is the only approach that ought to be acceptable from a scientific perspective. The subject matter of emotion theory is only vaguely delineated by folk discourse. That discourse tells us that emotions are marked by a number of features — phenomenology, physiological arousal, cognitive states, behaviour, etc. Which of these features will turn out to be central to the theory of emotion we produce is not to be settled a priori. Viewed from this perspective, Darwin's theory becomes a theory of the evolution of emotions themselves, or at least, of many important elements of those emotions. Darwin's approach is particularly promising from a scientific perspective, as the kind terms it produces will have the status that comes from being embedded in a major theoretical enterprise — the theory of evolution.

In the first half of the twentieth century it was customary to dismiss Darwin's work on the evolution of emotional expressions on the grounds that his evidence was mostly anecdotal or gathered in methodological flawed experiments. The favoured research programs were those of the anthropologists and students of "Kinesics", most notably W La Barre (1947) and R.L Birdwhistell (1963), who hoped to give an account of expression as a culturally specific, learnt code. With the revival of interest in evolutionary accounts in the past twenty five years, however, many of Darwin's results have been confirmed and extended. The research which has been the most compelling has been that on the universality of facial expressions in humans. In the next section I shall give a summary of this research. Having thus
demonstrated the correctness of Darwin's empirical data, I go on to discuss in section three the theoretical principles which he derived from it, and to assess to what extent these still seem plausible. In the rest of the chapter I outline the modern "affect-program" theory of evolved emotional responses, pointing out the similarities between this work and the modular approach to psychology commended by Fodor and others. I discuss some attempts to localise affect-programs in the brain, and in the final section, consider the limitations of affect-program theory as an overall approach to emotion.

3.2 Modern Work on Emotional Expression.

The recent experiments I shall review can be divided into four classes, judgment tests between literate cultures; judgment tests between literate and preliterate cultures; component analyses of spontaneous facial expressions; and component analyses of the expressions of blind born infants.

Izard (1969) developed a set of 32 photographs and presented them to observers in eight different cultures. His observers made judgments using a selection of emotion category labels which had been translated into their own language. Izard obtained convincing results showing very high levels of mutual recognition of his photographs but there were certain discrepancies. His results were weaker for Africans and Japanese than for other culture groups. Izard's African subjects were people from many parts of Africa all presently living in France and Izard believed that there were severe translation problems. Izard's Japanese translations have also been criticised by other translators and this may well account for his deviant results in that case. A similar experiment was conducted by Ekman and Friesen on observers in five literate cultures and they obtained no deviant results. But any data from cultures subject to western influence can be interpreted as merely showing that shared visual input across cultures allows members of different cultures to
recognise each others' facial expression conventions. In order to really demonstrate the universality of emotional expressions it is necessary to conduct studies on visually isolated cultures.

Ekman, Sorensen and Friesen (1969) conducted studies on the recognition of emotion in two visually isolated preliterate cultures, Borneo and New Guinea. They used the same stimuli from which positive results were obtained in five literate cultures. While the results were similar to the results from literate cultures the data were weaker and difficulties were encountered with the judgment test. It is difficult to show an observer a photograph and ask him to choose an emotional label from a list when, being illiterate, he is forced to memorise the list on each occasion. Their results were also criticised because the visual isolation of the tribes with which they worked was arguably incomplete. In their next experiment, Ekman and Friesen (1971) utilised a judgment task designed for working with children. The observer was given three photographs at once, each showing a face, and told a story which was designed to involve only one emotion. This method also has the advantage of avoiding the necessity of translating delicate emotion terms. The observers were from the Fore language group in New Guinea. They had seen no movies or magazines, they neither spoke nor understood English or pidgin, they had not lived in any western settlement or government town, and they had never worked for a white man. Forty photographs were used in experiments with 189 adults and 130 children from the Fore. Once again high degrees of agreement were observed between the categories which the pictures were intended to represent by the western experimenters and the categories which they were chosen as representing by the Fore. For example, in one experiment photographs intended by the western experimenters to represent sadness, anger and surprise were shown to the New Guineans who were asked to select which picture would represent the face of the man whose child had died. Seventy nine per cent of adults and eighty one per cent
Chapter three: The Psychoevolutionary Theory

of children selected the face intended to represent sadness\(^1\). These results confirm that the similar facial behaviour is selected for the same emotion concept in visually isolated preliterate cultures. Ekman and Friesen also conducted a reverse test using their New Guinean subjects. They asked visually isolated members of this culture to show how their face would appear if they were the person described in one of the emotion stories that had been used in the judgment test. Video tapes of nine New Guineans were shown to thirty four U.S. college students. Except for the poses of fear and surprise, which the New Guineans had difficulty in discriminating, the students accurately judged the emotion intended by their poses\(^2\). This reverse test further confirms the results of the study. The New Guinean end of the experiment was repeated by another team at a later date, with no discrepant results.

Ekman, Friesen and Malmstrom also conducted a component analysis of spontaneous facial expression. Averill, Lazarus and Opton (1969) had previously conducted research in Japan to establish that the eliciting procedure of having subjects watch certain stress inducing motion picture films led to self reports of similar emotions in both Japanese and United States subjects. In the Ekman, Friesen and Malmstrom experiment twenty five subjects from the University of Berkeley and twenty five subjects from the Waseda University, Tokyo each individually watched a neutral and a stress inducing film while alone in a room. They were aware that skin conductance and heart rate measures were being taken but unaware that a video tape of their facial behaviour was being made. When the repertoire of facial behaviours shown during the stress phase by the two sets of subjects was compared it was discovered to be very similar. Correlations between the facial behaviour shown by Japanese and American subjects in relation to the

\(^1\) Full figures in Appendix I
\(^2\) Full figures in Appendix I
stress film ranged from .72 to .96, depending upon whether a particular facial area was compared or the movement of the entire face.\(^1\)

Perhaps the most interesting result from the experiment was that when an experimenter was introduced into the room and allowed to ask questions about the subject's emotions as the stress film was shown again, the facial behaviour of the Japanese differed radically from that of the Americans. The characteristic difference was that the Japanese showed more positive emotion than the Americans and less negative emotion. The Japanese appeared to have masked their negative feelings by politely smiling. Slow motion video tape analysis showed the micro-momentary occurrence of characteristic negative emotional expressions, and then showed them being replaced with a polite smile. This behaviour appears to have been unconscious and relatively automatic on the part of the Japanese.

Finally, there have been component analyses of the deaf and blind born. Irenaus Eibl-Eibesfeldt (1973) examined the facial movements of infants born deaf and blind. He discovered that the same patterns of muscular activity were used to display the same kinds of emotions as in normal children. The possibility of tactile learning was also excluded in some cases, as some of the thalidomide children also lacked suitable limb structures. Eibl-Eibesfeldt's findings confirm the results of J Thomson (1941) and J.S Fulcher (1942), both of whom found the same patterns of muscular activity in the blind and seeing for each type of emotional behaviour. This result seems to have been fairly well confirmed and has extremely important implications for the theory of emotion.

Universality and innateness

Although these basic human expressions appear to be universal it cannot simply be inferred that they are innate. A universal response may be explained by various

\(^1\) Details of FAST technique for facial muscle movement analysis, and full results are given in Appendix I.
combinations of common learning and common innate inheritance. Consider for example the brow raising response as an expression of surprise. A strongly innatist theory might postulate the existence of genetic material which determines that the child will operate certain muscles in response to a strong stimulus through any of its sensory modalities, thus completely explaining the response. A less strongly innatist theory, on the other hand, might postulate that what is genetically coded for is only opening the eyes in response to visual shock, whereas opening the eyes in response to shocks through the other sensory modalities must be learnt as a way of acquiring more information about the source of those shocks. A strongly anti-innatist theory might postulate that all that is coded for is attempting to acquire more information when shocked and that the child learns that opening its eyes widely increases its visual field and thus the amount of information it receives. All these theories are prima facie plausible. Brow raising is a sufficiently basic and adaptive response to be plausibly innate. On the other hand there is a sufficient basis of common learning to make it plausible that such a response could be universal without an innate basis. There are experiments we could perform which would settle this dispute, but most would involve environmental deprivation of human infants, which is unlikely to prove acceptable to experimental ethics committees. There are, however, occasional tragic "experiments of nature", such as Eibl-Eibesfeldt's blind born infants, and phylogenetic continuity makes results gained with infant primates pertinent.

3.3 Darwin's Theory of Emotion Evolution

Darwin believed that the facts he had discovered about the expression of emotions could be explained by three general principles. The first of these Darwin calls the principle of serviceable associated habits. This states that a habit found serviceable to an animal in connection with some emotion that it has experienced in the past is
likely to be displayed in future when it experiences that emotion, even if it is no longer of service. Darwin's second principle is the principle of antithesis. This states that as certain emotions lead to certain habitual actions as mentioned in the first principle, when a directly opposite state of mind is induced there is a tendency to perform actions of a directly opposite nature. Darwin's third principle is the principle of action due directly to the nature of the nervous system. It is difficult to give a fair account of this last principle without running to some length. In this section I want to consider how Darwin's three principles hold up in the light of modern research.

The Principle of Serviceable Associated Habits

This is undoubtedly the least contentious of the three principles. Darwin renders it unnecessarily contentious by stating it in a highly Lamarckian fashion. Often it is not at all clear whether he is discussing the acquisition of a habit by an individual or the acquisition and inheritance of a habit by a species. He introduces his principle in the following way:

"certain complex actions are of indirect service under certain states of the mind, in order to relieve or gratify certain sensations, desires, etc; and whenever the same state of mind is induced, however feebly, there is a tendency through the force of habit and association for the same movements to be performed, though they may not then be of the least use." (Darwin 1872 P28)

We may see Darwin's Lamarckianism even more clearly in the following passage:

"the fact of the gestures being now innate would be no valid objection to the belief that they were at first intentional; for if practised over many generations they would probably at last be inherited." (Darwin 1872 P61)
But it is possible without too much difficulty to reconstruct Darwin's principle in such a way that it involves only the mechanisms of natural selection. Consider for example the baring of the teeth as part of the primate anger response. It is plausible to suggest that this behaviour was originally a preparation for attack. As such it may have had a direct survival value. The baring of the teeth in humans, however, is no longer a preparation to use them for fighting. At some stage during the development of the species this behaviour has lost its original function. But this is not to say that it is now useless. It has acquired an alternative role, namely that of signalling aggression to other members of the species. Its survival value in this new role may be at least as great as that it possessed in its old one. We have here a good example of a non-Lamarckian version of Darwin's principle of the retention of serviceable associated habits. A behaviour originally associated with an emotion because it fulfills some function will not cease to be associated with the emotion simply because it has ceased to serve that function. If it has a signal value, because of its past association with that emotion, it may become a permanent part of the species innate behavioural repertoire. In this reconstructed version, the principle is of great service in explaining the evolution of emotional behaviours which would otherwise seem quite arbitrary. Behaviours that originally evolved because they performed some utilitarian function acquire a signal function in virtue of their association with certain states of mind, and are retained in virtue of their signal function when they cease to perform their original function. Some emotional behaviours still have both signal and utilitarian functions, notably brow raising in surprise, which increases the visual field.

The Principle of Antithesis

Darwin's second general principle of expression is the principle of antithesis. Darwin's reasons for proposing the principle are purely empirical. He notes, for example, that:
"When a dog approaches a strange dog or a man in a savage or hostile frame of mind he walks upright and very stiffly; his head is lightly raised, or not much lowered; the tail is held erect and quite rigid; the hairs bristle, especially along the neck and back; the pricked ears are directed forwards, and the eyes have a fixed stare .... These actions .... follow from the dog's intention to attack his enemy" (Darwin 1872 p50-51.)

Darwin then considers the dog's response when he recognises the "stranger" as his master:

"let it be observed how completely and instantaneously his whole bearing is reversed. Instead of walking upright, the body sinks downwards or even crouches, and is thrown into flexuous movements; his tail, instead of being held stiff and upright is lowered and wagged from side to side; his hair instantly becomes smooth; his ears are depressed and drawn backwards, but not closely to the head. From the drawing back of the ears, the eyelids become elongated, and the eyes no longer appear round and staring .... Not one of the movements, so clearly expressive of affection, are of the least direct service to the animal. They are explicable, as far as I can see, solely from their being in complete opposition or antithesis to the attitude expressive of anger". (Darwin 1872 p51.)

Darwin's idea is that emotions can be seen as standing in relations of opposition, and that an animal disposed to act in one fashion under the influence of one emotion will act in the opposite fashion when under the influence of the opposite emotion. But it is not at all obvious that emotions stand in opposed pairs in the way that Darwin and many later theorists believe. There are three possible bases for the notion of "polar oppositions" between emotions. The first is the quality of experience that accompanies the emotion. Since Darwin conceives the emotion itself to be a feeling, as we saw above, I suspect that it is this first possibility that he has
in mind. One emotion just feels opposite to another, and this triggers an opposite set of behaviour. But this explanation of opposition would leave us without any etiological story. It would not explain why the animal has evolved to behave, or for that matter to feel, in this way.

The second basis for opposition is the action tendency associated with an emotion. In the case of fear this is flight, and in the case of anger, attack. These may, therefore, be regarded as opposite emotions because they lead to opposite actions. This explanation is not available to Darwin, as he wishes to use his notion of opposition to explain the occurrence of opposite action tendencies and this would render his account circular.

The final basis for classing emotions into opposites is the functional significance to the organism of the state of affairs to which they relate. Thus, fear is a response to dangers to be fled from, whereas anger is a response to challenges that are to be attacked. I show in later sections that these functional classes — challenges, surprises, dangers, etc — are the categories with which the environment must be divided up if we are to understand the psychoevolved emotions. Hence, using these classes as the basis for oppositions among emotions seems appropriate. Our reconstruction of Darwin's principle will run as follows. If a given emotion is associated with a class of states of affairs, such as challenges, and this leads an animal to act in a particular way, then an emotion associated with an opposite class, such as objects of affection, will lead to directly opposite actions.

The evolutionary rationale for this principle can be derived from Darwin's own comments. On page 61 he notes that:

---

1 Some philosophers would simply reject the use of sensations in this kind of causal role. I take a somewhat different view in Appendix III. I accept that a state which is a sensation might be involved in causing behaviour, and that it's relational properties vis à vis other sensations might be critical in determining its causal role.
"As the power of intercommunication is certainly of high service to many animals, there is no a priori improbability in the suggestion, that gestures manifestly of an opposite nature to those by which certain feelings are already expressed, should at first have been voluntarily employed under the influence of an opposite state of feeling". (Darwin 1872)

Darwin goes on to reject this view because he is unwilling to attribute such communicative intentions to animals. But we can strip his insight of the suggestion of communicative intent and merely note the survival value of clearly communicating emotional states and their opposites among animals of the same species. For a social animal, the display of an absence of aggression is as important as the display of aggression, since it placates more senior members of the group. An animal which clearly displays passivity by reversing normal signs of aggression has a survival advantage. This principle holds generally. Displaying antithesis is an effective mode of emotion communication and emotion communication in a social animal is often of considerable survival value. This view of the principle of antithesis will allow it to retain a place in the psychoevolutionary explanation of emotional expression.

The Principle of Direct Action

Darwin's third principle is the principle of the direct action of the nervous system. Darwin has an almost hydraulic conception of "nerve force" or "nerve energy", comparable to Freud's conception of psychic energy. In Darwin's conception, once a certain amount of nerve force has been released it is like pressurised hydraulic fluid and must find an escape valve somewhere - a means of expression in action. Thus he says:

"when the sensorium is strongly excited nerve force is generated in excess, and is transmitted in certain directions, dependent on the connection of the
nerve-cells, and, as far as the muscular system is concerned, on the nature of movements which have been habitually practised". (Darwin 1872 P66)

and again:

"good instances are afforded by the trembling of the muscles, the sweating of the skin, the modified secretions of the alimentary canal and glands, under various emotions and sensations" (P81)

Darwin appears to conceive of these responses as largely nonfunctional. It is this aspect of his account which is most likely to be questioned today. We now have a deeper understanding of the function of such responses as trembling, sweating and internal secretions. Trembling and sweating function, among other things, to control body temperature. The secretions of the various glands have important functions in preparing the body for various species of action. Because of this we will be less inclined to treat these as the arbitrary side effects of the changes in the central nervous system precipitated by the onset of emotion. Instead we will want to integrate them into our account of the functional, adaptive response that constitutes the emotion. In a modern psychoevolutionary account, autonomic nervous system responses should, hopefully, take their place alongside facial behaviour, action tendencies and subjective feelings as aspects of an overall integrated emotional response. The extent of the empirical evidence for or against the adaptation of autonomic nervous system response to the purposes of the various emotions is a vexed question, and is discussed at some length in section 3.4.

We are now in a position to summarise the role that Darwin's three principles are likely to play in a modern psychoevolutionary theory of emotion. The principle of the retention of serviceable habits, in its non-Lamarckian version, is likely to retain a central position in any theory. Our reading of the principle, however, will lay more stress on the direct survival value of emotional communication. We will see emotional expressions produced by the principle of serviceable associated habits not
merely as atrophied functional behaviours which happen to be expressive, but as behaviours which have acquired an alternative function and have been selected for on the basis of that function. The modern version of the principle of antithesis will also lay more stress on the survival value of communication. As we saw above it is the importance of clearly communicating emotions among the members of the species that gives an evolutionary rationale to the principle of antithesis. The principle of the direct action of the nervous system appears to be the most outdated of Darwin's three principles. We are now more inclined to see autonomic nervous system changes as functional and adaptive. A theory which treats the majority of these autonomic nervous system changes as side effects explicable only as brute facts about the physical structure of the nervous system will seem pretty unsatisfactory.


The evidence we reviewed in section 3.2 has led a number of theorists to revive the psychoevolutionary theory of emotions. The data has usually been handled within the framework of an affect-program theory of emotional response. The idea of an affect program and the psychoevolutionary approach have therefore come to be closely linked. An affect-program is a neural program which coordinates the various elements in the response. These are usually taken to include expressive facial changes, skeletal/muscular responses such as flinching and orienting, expressive vocal changes and autonomic nervous system changes. Some authors would add subjective sensations to the list.

When discussing the idea of an affect-program it is important to distinguish the question of whether or not affect programs exist from the question of whether or not they are innate. In principle an affect-program may be entirely innate, may be
entirely learnt or may consist of an innate core supplemented by learnt elements. These are questions about how the circuitry gets built, not about whether it exists. However, if learnt behaviour is to be considered as stemming from an affect program it is essential that it be automated. By this, we mean that it must proceed in an effective and co-ordinated fashion but without the need for conscious direction and co-ordination.

Although the existence of an affect-program does not allow us to assume that the response it co-ordinates is innate, there is an argument in the opposite direction, from evidence for the innateness of elements in the response to the existence of affect programs. If we believe that a co-ordinated set of responses involving several different bodily systems occur as a result of an innate factor, our present scientific understanding makes a programme-like structure in the central nervous system just about the only mechanism that could mediate between the genes and the production of such behaviour sequences. I shall therefore describe the evidence for thinking the various elements in the emotional response to be innate before offering direct arguments for the existence of affect-programs, and spelling out in more detail the precise nature of those programs.

Innateness is a difficult concept and all I can do here is to mention some particular points about it that are relevant to the argument that follows. First, in the rest of this section I shall be discussing the innateness of the response itself, not the innateness of what elicits the response. A response may be the innately specified expression of a given inner state, even if that inner state is not innately sensitive to any particular stimulus. The experiments of Ekman and others have sought to show that people in all cultures respond in a similar way to things that frighten them, not that people in all cultures are frightened of the same things. Secondly, in what follows I shall treat innateness in a gradualist fashion. A response is not innate or acquired simpliciter, rather innateness should be conceived as the degree to which the emergence of the response is insensitive to environmental changes.
The fact that a response has to be learnt does not exclude the possibility that innate factors influence that response. Suppose that a species is able to acquire some kinds of response without difficulty, but others only with difficulty. This sort of "learning bias" would implicate innate factors in learning. This possibility has long been admitted in one form, under the name of "triggering". In this case, a behaviour emerges after contact with some sort of environmental event, but it seems inappropriate to say that the response was learnt from that event. The duckling's imprinting on the first moving object it sees is a case of a triggered response. The response would be a highly appropriate to a mother duck. But the response does not emerge when there is good evidence that a mother duck is present. Instead, it emerges in conditions which, in the normal environment, happen to be coincident with the presence of the mother duck. In themselves, however, they are ludicrously insufficient evidence for the presence of the mother duck, as we see when we remove the duckling from its normal environment. Learning biases are less extreme instances of this phenomenon. A response appropriate to $p$ emerges when the evidence only slightly probablisies $p$, or a response appropriate to $q$ refuses to emerge unless there is overwhelming evidence that $q$. Learning biases are the preadaptation of the learning mechanism to the environment. Triggering is the limiting case of a learning bias. Conversely, the fact that a particular pattern of response emerges because of innate factors does not mean that it cannot be eliminated or reduced by learning. The blink reflex would be an example. To cope with these and other issues we need to replace the innate/acquired dichotomy with a continuum. Many authors have called this continuum the degree of "environmental buffering".

In the particular context of our discussion, it is reasonably clear what constitutes evidence of a high degree of environmental buffering, or of innateness. If a response occurs in all, or in a large majority, of members of a species, whatever their learning history, then this is good prima facie evidence. It is particularly good
evidence if the response in question is an arbitrary response, that is, one of a large class of responses, any one of which could perform precisely the same functions. Clearly facial expressions which serve to communicate emotions among members of our species are arbitrary responses. Only the surprise response retains much trace of its original, utilitarian function. It will be difficult to explain the occurrence of these in cultures which have no contact with one another, except by reference to innate factors. Darwin himself puts this neatly on page 15:

"Whenever the same movements of the features or body express the same emotions in the several distinct races of man, we may infer with much probability that such expressions are true ones - that is, are innate or instinctive. Conventional expressions, or gestures, acquired by the individual during early life, would probably have differed in the different races, in the same manner as do their languages..." (Darwin 1872)

Very early emergence of a display is also evidence for its innateness. An innate display need not be present in infancy, since it may require an environmental trigger. But if a response is present in very early infancy then it is likely to be innate, because of the relatively impoverished body of experiences from which it could have been learnt.

Reviewing the Evidence for Innateness

We have already presented a large quantity of evidence for the innateness of human facial displays of emotion. They have been shown to display a striking invariance across human cultures, and to emerge very early in life even in infants who are born without sight, and so have no means of acquiring those expressions. In Chapter seven we will present further evidence on the extraordinarily early emergence of sensitivity to facial expression in human infants, a sensitivity which strongly suggests that there are innate mechanisms for the recognition of these expressions, as well as for their performance. There is, therefore, a very strong case for
supposing human facial expressions associated with the basic emotions to be heavily influenced by genetic factors.

The next most heavily researched area of response is the activity of the autonomic nervous system. This research has been directed more to discovering whether ANS response differentiates between different emotions than whether it is innately specified, but the two questions are not unconnected. There are two reasons why a psychoevolutionary theory would find it more congenial if ANS response turned out to differentiate amongst emotions. Firstly, the added complexity this would give to the emotional response would be one more reason to posit affect programs to coordinate it. Secondly, and more importantly, if we think of ANS arousal as part of an adaptive response, we would expect it to reflect the forms of adaptive activity which the emotion is supposed to lead to. Thus, for example, the ANS arousal associated with fear and anger might be similar, since fighting and flying require much the same physiological preparation, but it would be odd if grief led to the same sort of preparations, since in grief they could serve no useful function.

The empirical research on ANS arousal differentiation is still inconclusive, but this is not for want of experimentation. There has been a vast amount of research on the topic. One reason for this has been the desire to test the famous James-Lange theory of emotion, first put forward in the 1890's. According to the James-Lange theory a conscious emotion is the perception of autonomic nervous system changes which follow directly (by a reflex arc) upon the perception of some external stimuli. The causal sequence which brings about our consciousness of fear, for example, is as follows. First, we perceive a fearful object. This precipitates autonomic nervous system changes as part of the flight response. The perception of these changes is what constitutes our consciousness of emotion.

This theory stimulated a long and productive tradition of experiments upon the autonomic nervous system responses involved in emotion. The first, and among the
most famous of these were conducted by Walter D Cannon in the 1920's. Cannon offered five experimental results which were supposed to tell against the James-Lange theory. First, he pointed out that cutting the spinal cord and vagus nerves in dogs does not appear to inhibit their ordinary emotional reactions. They still exhibit anger, rage and pleasure in handling. Secondly, Cannon presented evidence to show that autonomic nervous system responses are largely undifferentiated. Any stressful stimuli produces roughly the same set of responses. These include increase in heart rate, increase in blood sugar, increase in adrenalin levels, the erection of body hair and decreases in the secretions from the digestive glands and in the muscular activity of the alimentary canal. Thirdly, Cannon pointed out that the viscera are relatively insensitive structures and most of the time we are not aware of any sensation from that source. Fourthly, Cannon noted that the response time of the visceral organs and their associated nerves is relatively slow and appears to be too slow to account for the immediate response that occurs in the case of such emotions as fear and startle. Finally, Cannon pointed out that attempts to artificially induce the appropriate visceral changes by, for example, injecting adrenalin did not appear to produce the emotional responses that the theory would predict. At the time Cannon's work was thought to have killed off the theory once and for all. But in fact the theory has left as its inheritance a continuing tradition of experiment on the autonomic nervous system.

One of the most significant of these experiments in terms of its influence on emotion theory was that of S Schachter and J.E Singer (1962), which reopened the debate over the precise role of cognition in emotion. Schachter and Singer were interested in the results of G Maranon (1924) who had experimented with the effects of adrenalin injections. On the basis of their interpretation of his results they put forward three hypotheses. First, a subject will label a state of autonomic nervous system arousal for which they have no immediately available explanation in terms of the cognitions available to them at the time. Which emotion they report will
depend not on the type of arousal but on the cognitions available when they experienced the arousal. Secondly, if the subjects are offered an immediate physiological explanation of their arousal they will not feel the need to provide the arousal with cognitive labels. Thirdly, Schachter and Singer hypothesised that an individual would consider his cognitions to be emotions only to the extent to which he or she was physiologically aroused. Given the long and widespread influence exerted by Schachter & Singer's results, which are frequently cited by cognitivists to prove the irrelevance of physiological investigations of the emotions, it is worth considering their experiment in some detail.

Schachter & Singer divided their subjects into four groups. One group were injected with a placebo as a control condition. Of those injected with adrenalin, one group were told the genuine physiological effects that they would experience, another group was told nothing and a third group was misinformed about what they would experience. Each of these groups was divided again and was subjected to two different conditions. One section of the group was subjected to conditions designed to produce happiness or euphoria and the other section to conditions designed to produce anger. These emotions were to be induced by the behaviour by stooges placed with the subject and, in the later case, by the use of annoying and impertinent questionnaires. Schachter & Singer gathered results by making secret observations of their subjects while they were being subjected to the anger and euphoria conditions and also by asking their subjects to fill in questionnaires after the event. The results were not particularly clear. Robert Plutchik and other theorists have raised questions over Schachter & and Singer's interpretation of them (Plutchik and Ax 1967) But in general, the results were capable of an interpretation friendly to Schachter & Singer's hypotheses. The group who were fully informed about the effects of the adrenalin injection showed and reported the least signs of emotional arousal. The group misinformed about the effects of the injection showed
and reported the greatest degree of arousal and the group told nothing fell somewhere between the two.

Despite the tentativeness of these results Schachter & Singer's experiments have often been cited as if they were conclusive proof of the "cognitive labeling" theory of emotional arousal. A recent experiment, however, has produced quite contrary results. Paul Ekman (1983) used a more sensitive form of experiment in which a number of different indicators of autonomic system arousal were measured and note was taken of the patterns of relative change of these several different measures in the hope of displaying characteristic patterns for several different emotions. Ekman measured heart rate, electrical resistance of the skin, skin temperature and muscular tension. Ekman's results appear to show characteristic patterns for several emotions. Not only was he able to distinguish the so-called positive emotions, such as joy, from the so-called negative emotions such as fear, anger and grief, he was also able to distinguish among those negative emotions. The differentiation of autonomic nervous system arousal between the several emotions is thus clearly an empirical question still open to debate.

I have said that evolutionary considerations suggest that ANS arousal will be differentiated, but differentiation and innateness are not very tightly linked. ANS responses might be learnt even if they were differentiated. A child might watch its parents and siblings going pale and tense, and by trying to match these external manifestations, modify its own habitual vasodillation and muscle tension responses to fit the particular emotional situation. It is harder to get up a story about visceral changes, or changes in electrical skin resistance, but perhaps these invisible changes are physiologically linked to the visible ones, which would act like the monitor on biofeedback machines used to teach control of heartrate, etc. Whatever the outcome of the differentiation dispute, further experimentation will be needed on the innateness question. The natural methods, as in the case of facial expressions, are studies on infants and cross-cultural studies.
When we turn to other elements in the emotional response, such as skeletal/muscular and vocal responses we are forced to remain agnostic as few studies seem to have been done.

Since only the evidence on facial expression is conclusive at this stage, our conclusion about the innateness of the affect programs must be that the affect program clearly has an innate core, but that the boundary between that core and any learnt elements remains to be drawn. Nevertheless, it follows that if we accept the arguments for the existence of affect programmes we will be considering them as at least partially innate structures.

3.5. The Modern Psychoevolutionary Theory - The Case for Affect Programs

The main argument for the existence of affect programs is an argument to the best explanation. There are a number of important characteristics of emotional responses that seem to stand in need of explanation. They are often brief, quick, complex, organized and to a greater or lesser extent involuntary. Let us consider these characteristics in a little more detail. If we distinguish emotions from moods they are typically brief, they may last as little as a few seconds. The onset of an emotion may be so quick as to be regarded as near instantaneous. An emotional response is both complex and organized, it typically involves at least four kinds of co-ordinated changes. First, there are skeletal/muscular changes such as flinching, orienting or relaxing. Secondly, there are changes in the facial musculature, leading to emotional expressions. Thirdly, there are vocalisations and more general changes in the vocal muscles leading to expressive tones of voice. Fourthly, there are autonomic nervous system responses, such as sweating, secretion of adrenalin and changes in heart rate. We may also wish to mention the characteristic sensations of the various emotions. Perhaps the most important feature of emotional responses is that they typically occur without the conscious instigation of the owner. In this they are
rather like reflexes, but they are unlike reflexes in being complex and organized. The argument for the existence of affect programs is simply that the features of emotion which we have just described, and particularly the last feature, can best be explained on the hypothesis that a neural program stores a predetermined set of responses which are activated in a co-ordinated fashion in rapid response to some external stimuli. This argument has been advanced forcibly in a number of works by Paul Ekman (1972, 1973, 1980.)

The Automatic Appraisal Mechanism

Ekman goes on to point out that the affect program cannot respond to simple physical stimuli as the class of stimuli appropriate to a particular emotional response is usually a broad functional class such as novelty or danger. The input to the affect program must therefore be cognitively mediated. But the rapidity of the response, its reflex character, and particularly its informational encapsulation, that is to say its isolation from our other beliefs as in arachnophobia, mean that it is mediated by some independent system that is not wholly open to interference by our central cognitive processes. We should note that Ekman is not proposing that this mechanism is the only route by which affect programs can be triggered. For example, when we respond to an entirely novel stimuli which we assess as dangerous we may assume that our central cognitive processes are involved in triggering the affect program. But some special mechanism is needed to explain the way in which we respond with extreme rapidity, and apparently quite automatically to stimuli with which we are acquainted. Ekman calls it the Automatic Appraisal Mechanism. He conceives the mechanism as a memory which stores the lessons of our past experience and allows a more rapid response to important environmental events than would be possible if it were necessary to re-evaluate the significance of these events on each occasion. We shall consider further evolutionary reasons for the existence of such mechanisms below.
We have seen that the contents of the affect programs are at least partially innate. The same question about innateness arises for the appraisal mechanism. J.B Watson thought he had shown that newborn babies are afraid of loud sounds and loss of balance, but nothing else. But this makes no allowance for the possibility of innate responses in need of triggering. As we said in the last section, a response may be more or less innate. Experience may be needed to produce a response, but if less experience is needed to produce a response to one object than to another, the explanation of this predisposition, or bias, may invoke genetic factors.

Despite this reservation, however, it seems reasonable to suppose that the innate element in the appraisal mechanism is quite small. Broad a priori considerations from psychoevolutionary theory would lead us to expect this. The role of the mechanism is, after all, to adapt the affect program to changing environments and new classes of stimuli. It should be remembered that the appraisal mechanism appears to deal with stimuli at a fairly low level of perceptual analysis. Its role is to relate the broad functional class of danger, or noxious substances, to simple, readily recoverable, features of the environment. The functionally equivalent stimuli that are known to produce the same response in different cultures may have no such simple features in common. Children have to be afraid of bull ants, as well as electric sockets. It will not be possible to equip them innately to face such a wide range of environments. To accept this argument at face value would be to buy into a rather strong form of adaptationism, but given that we can't find many innate elicitors, it's nice to know that there's no reason to expect them.

**The Affect Program System as a Module**

The system of affect program and automatic appraisal mechanism that we have been describing possesses many of the features which Jerry Fodor (1983) lists as characteristic of modular cognitive systems. Its operation is mandatory. We do not choose to respond with fear or anger to a given stimuli. The system is largely
opaque to our central cognitive processes. We are aware of its outputs, which are
the emotion responses themselves, but not aware of the processes that lead to them.
Finally, and most importantly, the system is informationally encapsulated. There is
only a limited data base available to it on which to base its decisions as to the
appropriate response. Our ordinary background beliefs concerning, for example,
the harmlessness of earthworms, do not get taken into account when the system is
deciding upon a response. Informational encapsulation is one of the most important
features of modular systems, since it captures what we mean when we say that
modular processes are not part of general intelligence. In the case of the automatic
appraisal mechanism the module accepts perceptual information, presumably at
quite a basic level of analysis, and has access to our past learning history, in some
rapidly utilisable form, such as generalisations about the significance of certain
environmental features. Thus, when we are scared by a spider it is probably
something like, eight-legged things of a particular shape that trigger the response,
and a generalisation such as "anything that looks like that is dangerous" that
determines which affect program is run.

We said above that the appraisal mechanism is sensitive to our past learning history.
This is learning in its widest sense. We do not just recall actual experiences, but use
all our central cognitive processes to make assessments of classes of stimuli which
are then available to the appraisal mechanism. We can see this clearly in the case of
fear. What we store as dangerous seems to be sensitive, not just to experiences of
pain or unpleasantness, but also to observation of others' experiences and even to
observation of social and parental attitudes. Very few children have been hurt by
snakes, or spiders, or the dark, but it seems that only a slight demonstration of
revulsion or anxiety by adults is needed to produce a powerful aversion to these
things in the young. We can see how these assessments, once made, become the
subjects of far less intelligent processes from the way in which past assessments are
retained by the automatic appraisal mechanism when they are temporarily discarded.
Chapter three: The Psychoevolutionary Theory

by general intelligence. Sustained counter-conditioning appears to be needed to
delete an assessment once it has become linked to an affect-program response.

The Psychoevolutionary Advantages of Modules

Seeing the automatic appraisal/affect-programs system as a Fodorean module fits in
well with the general psychoevolutionary approach. Fodor, drawing on P Rozin
(1976) and others, gives a psychoevolutionary account of the reasons why modules
exist in systems such as our own which are capable of more general intelligence.
Modules have their origin in phylogenetic predecessors who did not have general
intelligence. They provide relatively unintelligent but effective ways of performing
certain low level cognitive processes such as perception, and, I am suggesting,
immediate emotive response. The question is therefore not why we have them but
why we have retained them. Fodor argues that the very unintelligence of modular
systems gives them a number of advantages over general intelligence. Among these
are the very short response times that can be obtained by having a mandatory
system and a limited data base. If a system is mandatory there is no decision time
incorporated in the response time. If the system operates on a limited data base,
which is chosen for its relevance to the question in hand, the procedures through
which it must go in order to make its decision may be more rapid. Finally, and most
interestingly, Fodor suggests that there are penalties to allowing our general
cognitive processes to interfere with our fairly low level processes. In the case of
perception he points out that it is vital for an organism to be able to accept data
which contradicts even its most centrally held beliefs. If perception is entirely a top
down process then the perception of novelty is impossible. The modularity of our
perceptual processes means that we are compelled to consider data hostile to our
present beliefs even if eventually we decide not to change them on the basis of that
data. Similarly, the modularity of our emotional responses can be seen as a
mechanism for saving us from our own intelligence by rapidly and involuntarily
initiating essential behaviours. Consider the startle response experienced when
confronted with novel stimuli. The affect program causes one to rapidly orient towards the stimuli, open one's eyes very widely and direct one's attention largely towards the stimuli. It also triggers preliminary elements of a flight or fight response in case this should be necessary when the stimulus has been evaluated.

Consider next a fear response. No matter what our higher cognitive processes tell us about the situation, if we have experienced an object as harmful in some past segment of our learning history, the AAM will trigger our fear affect program and this will initiate expressive facial changes and the necessary autonomic nervous system changes for a flight or fight response. In such a matter any false positives are distinctly preferable to false negatives! It is, of course, not necessary to our argument that any particular modular process, or even modular processes generally, are useful to us today. It is only necessary that in some previous state preserving our modular responses may have been more useful than incorporating those areas of our activity into general intelligence.

A Model of the Affect-program System

Figure 3.1 shows the essential structure of the affect program system. The diagram shows the causal path from perceptual input through automatic appraisal to affect program and finally to the series of changes which constitute an emotional response. The diagram also shows how perceptual input to central cognitive processing leads to initial evaluations of classes of stimuli, which evaluations are then stored in the automatic appraisal mechanism. It also shows the direct triggering of affect programs by central cognitive processing in cases where a normal stimuli is evaluated for the first time. Some such link is presumably also needed to account for cases where imagining emotionally arousing stimuli leads to a genuine emotional response. Finally, the diagram includes a possible pathway for conscious or unconscious intervention which would be required for the operation of display rules. A display rule is a means whereby an individual, or more often a whole society habitually mask or suppress a natural emotional response in order to
conform more closely to some norm or ideal. The classic example of display rules in action is provided by the Ekman and Friesen experiment on Japanese and American university students described above. The Japanese were shown to modify their natural emotional responses in the presence of authority figures while still retaining their tendency to respond by micromomentary expressions which occurred in the instance before a polite smile was imposed upon the face.

The theory we have outlined will complicate the relationship between folk psychological content attribution and emotions. When we attribute an emotion content such as "he is afraid of spiders", we could be attributing that content either to the automatic appraisal mechanism or to central cognitive processing or to both. If we are attributing it to central cognitive processing then it does not follow from our attribution that the subject will actually exhibit the symptoms of fear. If we attribute it only to the automatic appraisal mechanism it does not follow that the subject will admit to an appropriate belief such as "spiders are dangerous".

The ability to draw this distinction is a great advantage of the theory since it allows it to avoid the the paradoxes which have traditionally been created by cognitive theories. Objectless emotions occur when affect-program responses are inappropriately triggered. Reflex emotions occur when the automatic appraisal mechanism has contents which conflict with our central belief store. Unemotional evaluations occur because, for example, more than a belief in a future danger is required to trigger a real fear response. The modular, affect-program theory of emotions gives us what Stocker was looking for when he talked of merely "emotionally held" thoughts\(^1\). The "judgements" about the environment which are embodied in emotions are the judgements of a modular sub-system, not necessarily of the whole person. Having this distinction does not immediately solve such problems as the causation of emotions by imagined objects, but it reduces them to

\(^1\) See section 2.4
Chapter three: The Psychoevolutionary Theory

Figure 3.1. Model of Affect-Program system in a Fodorean Cognitive Architecture.
problems of substantial psychology, rather than philosophical paradoxes. The fact that we can have emotions without having appropriate central beliefs and desires ceases to be problematic.

3.6 The Neural Basis of the Affect Programs

What is known about the neural basis of affect program responses suggests that they are subserved by neural circuits in the limbic system, the phylogenetically old portion of the cortex which surrounds the brain stem, or in associated brainstem structures. The hypothalamus in particular is a plausible localisation for many of these responses. The central role of the hypothalamus in emotion has long been recognised. W.D Cannon's attack on the James/Lange theory in the 1920's was partly motivated by a desire to clear the ground for his own hypothalamic theory of emotion. Cannon and others had, even by that early date, amassed a large body of data on the neural basis of emotion.

MacLean's Theory

P.D. MacLean makes an important link between the idea of limbic localisation and the phylogeny of basic emotional responses (MacLean 1952, 1969, 1970, 1980). He attempts to explain the modes of functioning of the limbic system by its phylogenetic status as an inheritance from the lower mammals. MacLean is particularly struck by the reports of certain classes of epileptics. These patients report epileptic "auras", experiences which precede epileptic seizures. The seizures are related to lesions in certain areas of the limbic system, and the neural activity which accompanies the seizure is wholly confined to limbic areas, stressing the anatomical separation between this region and the rest of the cortex. The auras reported include emotional affects, such as those of fear and sadness, need related
affords, such as thirst, cold and the need to excrete, and stranger phenomena, such as déjà vu, and feelings of great certainty, of having grasped what is at the root of everything, which occur without reference to any propositional content. Some of these phenomena can be reproduced by the direct stimulation of brain sites. MacLean's interpretation of these reports is that the limbic system preserves a phylogenetically more primitive way of processing information and directing action than that subserved by the neocortex. As well as emotional functioning, he suggests that it may be involved in producing our irrational attachments to certain convictions, and our intuitive, and often fallible, sense of what is real or imaginary, remembered or invented, etc.

To a philosopher, this sounds like a return a philosophy of mind not sighted since Russell's "Analysis of Mind" in the 20's. But it would be a mistake to assimilate what MacLean says to traditional philosophical theories that stress the role of images and sensations in mental processes. MacLean is not proposing to analyse what it is to think, or know, or remember. As far as I can gather, he would give a relatively conventional account of these higher mental processes. All MacLean is aiming at is an account of certain subrational mental processes. He doesn't claim that what it is to be really convinced is to have a particular sensation. Instead, he observes that conviction is often reported to be accompanied by sensations, and tries to give an account of the neural basis of this phenomenon. We can accept MacLean's theory even if we then go on to give a thoroughly behaviouristic account of sensations, although I think there are independent reasons for not doing this, as I will argue in Appendix III.

MacLean's work is highly speculative, but there are good reasons for taking these speculations seriously. In section 3.4 we discussed Fodor's explanation of the existence of modular systems in the mind/brain. This involved the retention in our own, developed brains of ways of doing things originally adopted by our phylogenetic precursor for want of the resources to perform the functions any other
way, and retained because of their efficacy in various other respects. If we accept such arguments, a natural place to look for the localisation of those brain functions is in the parts of our brain equivalent to the brains of our precursors. It is by no means safe to assume that preservation of structure indicates preservation of function, but it does provide a rational basis for a research strategy, and evidence such as MacLean's which supports this approach should be taken seriously.

Panksepp's Theory

A recent attempt to locate neural circuits subserving basic emotional responses in other mammals is to be found in Panksepp(1982). Panksepp suggests that four pathways can be found in the hypothalamus which subserve emotional responses. Panksepp calls these the "expectancy", "fear", "rage" and "panic" pathways. His main technique for locating such pathways is direct stimulation of cells in the hypothalamus, evoking various sorts of behaviour. The aim is to trigger connected behaviors of the decorticate rage variety in order to demonstrate the existence of a hard wired basis for complex behaviour sequences which could be triggered relatively automatically by major life-challenges, such as danger or the presence of a potential fulfiller of a basic need. Panksepp's approach has met with some success, enough to lead him to propose his four pathways. He is also admirably conservative in refusing to posit any hypothalamic emotions for which he is unable to locate a pathway and specify a set of resultant behaviours. But I do have a number of reservations about his methodology.

Firstly, I think an approach which concentrates on relatively long term emotional responses, such as flight behaviour is likely to get the account of affect-programs wrong. It will ignore responses whose major manifestations are facial and other signal behaviour. It will thereby make it difficult to give a serious account of the commonalities between human and animal emotion, since it seems plausible that in humans only very short term muscular/skeletal responses, such as flinching, are
controlled by the affect program. Longer term responses seem to be mediated by higher cognitive processes. It would seem quite plausible to suggest that the growth of the neocortex is accompanied by the loss from the affect-program response of longer term elements in the emotional response, such as fighting and running away, and the transfer of these functions to more sophisticated and flexible cognitive mechanisms. This would constitute a major difference between emotions in humans and in most animals, and in consequence, an account like Panksepp's, which concentrates on longer term responses, would be less effective as a base on which to build an account of human emotions. The predominant feature of human responses seem to be the expressive element, but this was enough to suggest to Darwin a series of important similarities between humans and animals.

Secondly, I have a closely related worry about some limits Panksepp puts on the term "emotion". He does not care to investigate "surprise" and "disgust", although he admits that they are likely to be controlled by brainstem circuits. He thinks they are "too reflexive in character". Surprise and disgust do not seem any more "reflexive" in the most usual sense (e.g: mandatory) than fear and panic to me, so I presume that Panksepp is referring to their duration. Like reflexes, they seem to last only as long as the stimulus. They lack the "regenerative feedback" which he takes to be characteristic of emotions. I think Panksepp would find it easier to take this line with startle than with disgust, but in both cases, I think his stress on the relationship of emotion to longer-term behaviour sequences is narrowing his perspective too much. Clearly this is as much a terminological as a substantive dispute, but I do think that there is more interest in delimiting the class of hard-wired, hypothalamically localised responses to functionally significant stimuli than in delimiting the sub-class of those which sustain themselves for longer then the duration of the stimulus. My reason for thinking this is largely what I said in the last paragraph about emotion in humans.
Finally, and again connectedly, I think Panksepp's account would be improved if he was to look for the neural basis of the emotion categories revealed by ethological work, such as that of Darwin, Ekman and Eibl-Eibesfeldt, rather than the categories of a broad ecological theory of emotions, such as that of Plutchik (1962, 1980) or that which he himself describes. On the latter approach the preferred strategy is to develop a classification of universally significant classes of stimuli, such as dangers, challenges from conspecifics, social loss, fulfillers of needs, etc. Emotions are envisaged as responses designed to be adaptive in the face of one of these ecological classes. I agree that emotions are designed to be responses to stimuli with important ecological significance, but I disagree with the idea that we can decide what the significant classes are prior to examining the emotions themselves. It seems to me that our general ecological theory is in a pretty primitive condition. One way to improve it is to find out what emotions there are by direct ethological examination, and adjust the ecological theory accordingly. The experimental methods of the ethological approach are very sound, and are likely to give us a good taxonomy around which to structure our other research. I hope I've described them sufficiently above to justify this claim.

To summarise, the general order of research should be ethology, backed by neuroscience, and on the basis of this we can generate a broad ecological/evolutionary account of emotions. That, after all, would parallel nicely the generation of accounts of morphological evolution. First determine function ethologically, and develop a morphological classification, and then proceed to develop a broad theory of the dynamics of evolution and the phylogeny of the several species. Working the other way around would seem risky.
3.7 Limitations of the Affect Program Approach

We have now given a fairly detailed picture of the affect program theory of emotions and the evidence for it. Examples of emotions which do not fit the model spring immediately to mind. Love, in its romantic sense, is hardly going to be analysed as a rapid response which had survival value for our phylogenetic predecessors (though lust, of course, plausibly might be!) It is therefore important to understand that the affect program theory is not intended to be a theory of all those mental events which it is good English to call "emotions" or "emotional". The theory hopes to deal with a class of relatively simple emotions, many of them present in other animals, especially primates, as well as in the human species. These emotions exhibit to the greatest extent those properties which mark off emotions from such mental states as judgments and beliefs, features like passivity and insulation from our general belief store. Arguably, the theory deals with just those states that constitute our model for emotions and emotionality, the phenomenologically compelling, occurrent instances of anger, fear, joy, sadness, disgust and suprise. The fact that the categories of mental events explained by the theory are not identical with our pre-theoretic emotion categories should be of little significance beside the consilience of our theory with the rest of the biological and psychological sciences, and the very great explanatory power of the theory when compared with any other theory of emotion on offer. The psychoevolutionary theory explains why the emotions with which it deals should consist of both physical responses and assessments of the significance of the environment. The physical response is appropriate to the assessment of the significance of the environment and promotes the survival of the organism in that environment. The theory explains why emotions have the particular kinds of automation and modularity that are the distinguishing features of this class of mental events: the features traditionally referred to as the passivity of emotion. This is explained by the evolutionary advantages of modular emotional responses discussed above. The
Chapter three: The Psychoevolutionary Theory

theory explains why we have emotions. It does this by integrating them with the theory of evolution. It also gives us a new approach to the traditional questions about the rationality and moral status of these emotional responses, often making those questions seem as absurd as they would be if asked about the blink reflex or the processes of visual image construction. Attempts to answer those questions without a substantive scientific account of emotions come to seem willfully obtuse.

We should, I have said, be unworried by the introduction of these new, smaller categories of emotions. We should be equally unworried by the fact that whole emotion categories, such as love and envy are left for some other theory to deal with. To insist on a unified account of all the things English speakers call emotions is to be absurdly devoted to the prescientific categories of a particular phase of a particular culture. I shall be arguing at length in chapter five that it is such "univocalism" - the desire to keep "emotions" as a natural kind- that has bedeviled emotion theory, and caused so many absurdly procrustean views to gain currency.

In the rest of this section I want to consider one such attempt to get a univocal theory by forcing the affect program theory to fit the univocal bed. This is Leventhal's "Perceptual Motor Theory" (Leventhal 1984). I shall try to show that Leventhal's account is an unhelpful distortion of the affect-program approach.

Leventhal's proposed model is summarised diagrammatically in figure 3.3. His "emotion system" has three levels, each primarily responsive to certain classes of stimuli, although partly responsive to the adjacent types of stimuli, as shown by the dotted arrows. The levels also have characteristic classes of output. Leventhal hopes to give a model that will cope with all emotional phenomena by treating instances which depart from his preferred pattern as cases of only part of the emotion system operating. This will allow him to claim, for instance, that the
objectless fears that I cited against the cognitive theory are really only partial fears, and that they are really part of the same phenomena as purely cognitive fears.

Leventhal's "levels" deserve some detailed examination. Level three, the "expressive-motor" level is supposed to involve responses which are little more than complex reflexes. They are innately sensitive to certain triggers, and serve to explain the expressive behaviour of newborn infants, and the most universal and invariant features of adult responses. Level two, the schematic level, serves to explain how rigid, pancultural responses can be sensitive to past learning history.

---

1 "I'm afraid we're going to be a little late tonight dear".
Leventhal postulates a store of "schematic", "prototypical" memories, derived from experience, and becoming more abstract as the individual undergoes a wider range of similar learning experiences. These memories play the same role as the automatic appraisal mechanism in my earlier model of affect-program responses, except that innate responses are dealt with separately. Level one, the conceptual level, consists of abstract, propositionally stored information. Processing at level one is "voluntary and effortful, and makes demands on conscious attentional capacity." (1984 P:277)

As I've said, Leventhal hopes to use the three levels, in various combinations, to account for all standard cases of emotion. When John is frightened by an obviously harmless loud noise, it is level three processing that comes apart from one and two. When John is afraid of a rubber spider on our hand levels one and two function without level three, and when John is in the throes of romantic love for a girl he has never met, level three is primarily operative.

Leventhal's distinction between levels one and two seems to me to be quite unnecessary even given the desire for univocalism. It seems to be motivated by a desire to make his account congruent with a Piagetian account of the stages of children's cognitive development. Leventhal cites various other phenomena which are supposed to involve "schematic" emotional processing. Chief among these is phantom limb pain, and the tendency of that pain to reflect past pains in the limb prior to amputation. I find the notion of "schematic memory" deeply obscure. Leventhal's main grip on his concept seems to be via the idea that some memories are "non-verbal" or non-propositional. He sometimes seems to identify this with not being present to consciousness, but he also associates "schematic" memory with the storage of information in the form of images! There seem to me to be a number of possible distinctions in play here, and little reason to suppose that they converge to mark any one distinction between different ways of remembering.
There are also general reasons for not wanting to split up apparently non-cognitive emotional responses into two kinds. Most importantly, there is the need to avoid making too sharp a distinction between those stimuli which the emotion system is innately sensitive to and those whose significance must be learnt. I have suggested that innateness is best conceived as a matter of degree. This will allow us to accommodate data which suggests biases in the learning abilities of the emotion system, and conversely to allow the right degree of malleability in the responses which children are born with. Leventhal's two systems, one innate and one learnt, and his notions of qualitatively different kinds of thought/memory (as opposed to information storage and processing by different, informationally encapsulated systems) constitute a theoretical step backwards of no small magnitude. Those of us not antecedently attached to Piagetian doctrines will find this unappealing.

Let us now turn to Leventhal's level one, the conceptual or propositional level. This part of the system is characterised as a structure of propositionally encoded information deriving from reflection upon past experiences. Utilising this information makes demands on conscious attentional capacity. It is of the same type as the conceptual information storage and processing connected with unemotional action. The obvious question, therefore, is why we should regard it as part of the emotion processing system, rather than as the application of our general cognitive machinery to a particular subject matter? There is no good answer to this query, so from simple Occamism we should regard it as part of our general cognitive activity. Thus, Leventhal's "cognitive level" turns out to be equivalent to the claim stated above, that cognitive processes are capable of directly triggering affect program responses. The sole motivation for including these cognitive processes in a separate "emotion system" is to force the mind/brain to fit our pre-scientific categories.

There is one final objection to Leventhal's account which should be stressed. Leventhal thinks emotion categories are unified by the convergence of the various elements in emotional responses in a way that suggests a limited number of
discrete, underlying events. This gives him the same account of the taxonomy of emotions as Ekman and the other affect-program theorists. But this agreement is only superficial. For Leventhal, the full blown emotional response contains various cognitive elements, and the cases which originally drove the affect program account are regarded as deviant ones where only part of the emotion system is functioning. But there is no serious scientific taxonomy of these "full responses". The experimentally well individuated responses are precisely those that easily come apart from cognitive factors, Leventhal's deviant cases! All Leventhal has done is take the folk taxonomy, or some fairly stipulative restriction of it, and speculated that its categories will fit into some serious theoretical taxonomy, and that some real basis will be found in the brain for the functional structures he describes in figure two. This approach stands in sharp contrast to that of the original affect program theorists. They based their account of emotion taxonomy on the classes utilised in their best theory. Leventhal bases his account of an untested theory on a previously decided taxonomy!

Leventhal's project is a classic example of the effect of univocalism on emotion theory. A theory of some of what we've traditionally called emotions comes along, forcing us alter our traditional categories to better reflect what we now know about the phenomenon. The univocalist then appears, arguing that our theory is inadequate because it doesn't deal with all the data and wanting to supplement it with speculations whose only justification is that they preserve the old taxonomy. But in an area where the current taxonomy is not tied to any useful body of generalisations, taxonomy should be very largely driven by theory. Those events should be grouped together whose grouping best contributes to finding and stating the generalisations covering them. It is as if someone wanted to alter the phylogenetic classifications of species because whales are so clearly a kind of fish. If anything, it is worse, because there's a good deal less reason to group the startle
response with Dante's love for Beatrice together than there is a shark with a dolphin.

While it would be a mistake to attempt to push the affect-program theory beyond the limited class of emotions for which it was designed, this stricture does not necessarily apply to the general idea of an psychoevolutionary approach to emotions. In chapter five I discuss the prospects for an evolutionary account of certain cognitive emotions. This approach may broadly be described as sociobiological. Unfortunately, there has been insufficient detailed work on this topic to allow a real assessment of the prospects for such work, and our comments will have to be largely programmatic.
4.1. What is Constructionism?

The constructionist theory of emotions has found favour among psychologists, philosophers and anthropologists. In this chapter I shall concentrate on the work of the psychologist J.R Averill, and on the recent series of papers advocating philosophical constructionism by Rom Harre and Claire Armon-Jones. In the introductory section I try to give a general feel for the constructionist position, concentrating on each of these three authors in turn. Constructionism as expounded by these authors is a diverse and inchoate doctrine. My own version will be very different from any of theirs. In my presentation of these other authors views I aim to show that they have encumbered the central insights of constructionism with many false and confusing auxiliary theses.

In the remainder of the chapter I discuss some basic problems for constructionist approaches to emotion, and show how the constructionist approach can make a contribution to our overall understanding of emotional phenomena without running into these problems.

Averill's View

At the beginning of his 1980b, Averill lays down the central tenet of the constructionist (or, as he prefers, constructivist) faith:

"First, emotions are social constructions, not biological given." (Averill 1980b p305)

"Biologically given" seems to mean genetically coded, and I suspect that it is the psychoevolutionary theory to which Averill is throwing down the gauntlet. An
emotion is "constructed" if it is a learnt response, whose elements are dictated by the ideas prevalent in the surrounding culture, rather than by the genes. Averill, like most constructivists, concedes that there are pre-social elements in emotional responses, such as the startle response, but he denies that these are fully fledged emotions in themselves, or that the role of pre-social elements in fully fledged emotions is very significant:

"Biologically determined reactions form a relatively small class of emotional reactions. Their primary importance is that they may be incorporated as elements into other kinds of emotion." (Averill 1980a p39)

"most "natural" fears (e.g. of falling, and even of bears in woods) are not very representative of emotions in general....emotions as we know them are the products of reflective experience. The person who meets a bear while alone in the woods at night must improvise and interpret his reactions according to categories of thought....acquired during socialisation; and it is such improvisation and interpretation that allows the response properly to be called fear." (Averill 1980b p325)

But this exclusion of "natural fears" is clearly stipulative. Most people would be quite happy to call such an immediate reaction emotional, and would take evidence of "improvisation and interpretation" to show remarkable self-possession, rather than real emotion. Later, Averill admits that it would be "fatuous" to give a constructionist account of a person's response when a bear leaps at them, so unless we think the interests of theory construction are best served by stipulating that this isn't an emotion, we are going to have to see constructionism as a theory of only some emotions.

The words "improvisation", and "interpretation" occur frequently in Averill's theory, as in the above quotation. He gives the impression that they are part and parcel of the constructionist view, but this is not the case. Averill compares learning
emotions to learning a language. The individual acquires a vocabulary of responses, and a set of rules which embody a society's sense of "appropriateness". The rules tell them when to use the responses. Averill then tells us that a person with this system of constructed states will interpret stimuli as falling into socially significant classes, and improvise an appropriate response. But these processes, if they occur, are not essentially linked to the fact that the response is a constructed one. A response may be constructed without having to be improvised each time it occurs. A car driver's response to a roundabout is presumably constructed, and is far too complex to be dismissed as a mere "motor memory", or some such phrase. Nevertheless, it can, and should, be so inculcated as to be produced without conscious direction, and even without being consciously noticed. So in fact, a constructed response need not be "improvised".

It might be argued that, although not all constructed responses need be improvised, only constructed responses have to be improvised. A biologically determined response would, in some sense, be automatic. But this is not true unless we extend our use of "socially constructed" to cover all cases of learning! A response which happens to be improvised could have been learnt in all sorts of ways, or even invented on the spot. It certainly need not be a reflection of the shared values of a culture.

As for interpretation, the psychoevolutionary theory, which posits a hard-wired, inherited emotional response still insists on the interpretation of the stimuli. It is hard to think of a modern emotion theory that doesn't require at least a minimal cognitive interpretation of the stimulus\(^1\). Interpretation is only a necessary component of constructionism because it is a necessary component of any plausible account.

\(^1\) Older theories, such as William James' and perhaps J.B Watson's wouldn't have required interpretation.
The real status of "improvisation" and "interpretation" in constructionist theory is that interpretation is uncontentious, and improvisation is an additional hostage to fortune which could easily be surrendered. No-one should be convinced of the truth of constructionism because they are convinced of the necessity for interpretation of the stimulus or the improvised nature of emotional responses. Both, for example, are available to the cognitivist.

If we really want to get to the heart of constructionism it is the notion of social construction itself that we should focus on. Averill's talk of emotions as social roles is illuminating here:

"An emotion is a transitory social role (a socially constituted syndrome) that includes an individual's appraisal of the situation, and is interpreted as a passion rather than as an action" (Averill 1980b p312)

The notion of a syndrome need not detain us, Averill merely wants to stress that an emotion is a loose collection of many different elements, not all of which need be present for it to be manifested. The clause about appraisal is a nod to cognitivist theories, with their stress on the need to interpret the stimulus. It is the first and last clauses which are the heart of constructionism. A social role is a way of behaving which serves various ends, both individual and collective, in a society. The role of father, for example, is both a way for the individual to define his role in society, and aids the collective functioning of that society. The notion that an emotion is a (transitory) social role suggests that emotions are cultural artefacts, which we can really only understand when we see the ends which they serve in the overall social system. The idea of interpreting emotions as passions is essential to making the social role idea plausible. It is highly counterintuitive to say that emotions are things that we "put on", in the way that the phrase "social role" suggests. To get round this problem constructionists propose that our belief that emotions are biologically
primitive and beyond our control is a myth which helps the emotion to fulfill its various functions. Our helplessly "suffering" emotion is part of the role we take on:

"Take anger as a case in point. There is a general cultural prohibition against intentionally harming another; however under certain circumstances...retaliation may be expected, and even demanded; it must, however, be carried out in such a manner that the individual does not willingly violate the general cultural norm against injuring another. Being "overcome" by anger is one way of meeting this dual standard." (Averill 1980a:p66)

The distinction between having an emotion in this way and pretending to have an emotion is delicate, to say the least. We will see below that constructionism has some difficulty in giving an account of pretending, particularly as some constructivist theorists have an antipathy to positing unconscious processes. Nevertheless, in at least some cases, this idea of "disclaimed action" provides a very plausible account of social phenomena. Averill's most convincing example is that of the Gururumba of New Guinea, taken from Newman (1964). The Gururumba experience the phenomena of "being a wild pig". In this state they run wild, looting, and shooting arrows at bystanders. The Gururumba think the wild pig syndrome is caused by being bitten by the ghost of a recently dead member of the tribe, and believe that this releases impulses suppressed by society and civilisation. The syndrome is treated as a disease by the tribe, and cured by ritual. Wild pig behaviour is largely restricted to males between the ages of twenty five and thirty five. This age is associated with the loss of youthful freedoms, and the acquisition of a wife, and social responsibilities. The thought naturally springs to mind that these pressures are the real cause of the syndrome. Related syndromes are found in a number of south-east Asian societies, referred to as "amok". This consists in indiscriminate attacks on others, and usually culminates in the killing of
the person who runs amok. Amok is traditionally triggered by perceived dishonour. Cases are cited of westerners living in Asia running amok, presumably by example.

The important thing about these examples is that we seem to be able to understand how the behaviour exemplified by them can be learnt, and rationally motivated, without being normal intentional action. The person running amok is not pretending to be in a frenzy, he is in a frenzy. But it is a frenzy that he has good reasons for being in. Just what this sort of semi-intentionality could consist in I will discuss in a later section, but for now all we need to note is that the essence of the constructionist strategy is to suggest that emotions are semi-intentional actions of this sort.

Harre's View

So much for the psychologist's account. Turning to the philosophers, constructionism has been taken up enthusiastically by Rom Harre. Harre (1986) gives a summary of the sorts of evidence that motivates his view. He argues that wide cultural variation in emotions suggests that they are shaped by the social system, and militates against "naturalistic" theories. He gives examples of this "cultural variation among emotion systems" under five heads (1986 pp10-12)

First, the moral connotations of emotion words may be different. Harre cites Lutz's (1986) example that "fear" is condemned in our culture, but "metagu", usually translated by "fear", is generally praised in the Ifaluk atoll culture. I find this particular example totally unconvincing. First, it is not at all clear that our moral views are so different. Perhaps the Ifaluk share the values of certain modern feminists, or, when they praise the fear of spirits, they are like those Christians who praise the fear of God. But suppose we give Harre his first point. He then goes on to say that:
"The moral qualities with which the concept is associated are such that it would surely be a gross mistranslation to treat metagu as equivalent to "fear"."

If this is the case, what is the basis for saying that we have one emotion with different moral values? Harre can't have it both ways. Presumably, Harre is trying to say that metagu has much in common with fear, but lacks its moral overtones. Since Harre thinks that moral overtones are central to understanding emotions, it follows that we shouldn't translate metagu by "fear". I find this essentialist view utterly uncompelling, as, clearly, did Lutz, who translates "metagu" as "fear/anxiety" throughout her article.

Secondly, Harre points out that some emotions are encouraged in one culture, but not in others. This is presumably very much like the last point.

Thirdly, a strong form of an emotion can exist in one culture, whereas only a weak form exists in another. Spaniards, he alleges, more readily and acutely feel embarrassment at the social ineptitude of others. They have a special, albeit phrasal, name for it. This point relates to Heelas'(1986) notion of "hypercognition". Emotions, says Heelas, may be hyper- or hypocognised by cultures. Roughly, this means that people talk about them more, or talk about them less. As a result, it is claimed, they classify more or less of their emotional life in those terms, and attribute and self-attribute those emotions more or less often. Love, for example, may be hypercognised in our culture when compared to some others.

Fourthly, claims Harre, emotions go in and out of existence in one culture. His example is "accidie", a form of depression/boredom found in hermits in early Christian times, and in medieval monks. Harre has a joint paper on this emotion later in the same volume. Unfortunately, his co-author, the clinical psychologist Robert Finlay-Jones, writes consistently as if accedie is a state found in many modern subjects, and nowadays not distinguished from depression. This doesn't fit
very well with the claim that emotions come and go with the flow to history. It suggests that what's going on here is hyper/hypocognition again.

Fifthly, Harre says that quasi-emotions, such as "cosiness" also exhibit cultural variation. The Dutch have a word which means something more like "with friends" which is used in similar contexts, and the Finns would use "homely" instead of cosy. I don't really see the point of this section, unless it's to show that it's not just emotion words that seem to cut up the phenomena differently in different cultures. But that seems a bit obvious.

I find Harre's summary of the evidence for cultural variation in emotions pretty unimpressive. It's supposed to support the thesis that the elements in emotions that are common to all cultures are trivially small. If anything, it seems to do the opposite. Even cultures in which states are "ultrahypocognised", as it were, seem to have some trace of them. Rather than disappearing, as they should do if wholly constituted by their role in the social structure, such emotions seem to be assimilated to others, creating broader categories, but remaining discernible to the anthropologist.

**Armon-Jones's View**

I think the main reason for Harre's and my very different responses to the evidence is that he doesn't deploy a distinction between an emotional response (e.g., what someone does/is when they have an emotion) and the stimuli that elicit it. This is equally apparent in the work of his colleague Claire Armon-Jones. Armon-Jones (1986b p64) takes the "naturalist" to be arguing that emotions are "universal and natural dispositions" to respond in a certain way to certain stimuli. Fear of the dark might be such a natural disposition. Constructionism supposedly has the advantage over "naturalism" of being able explain fear in situations that "do not naturally warrant fear"(p65). Armon-Jones' example at this point is a tribe whose children fear their elders and the neighbouring tribes. It is a problem for the "naturalist" that:
"while actions such as "fleeing" take the same form whether culturally or naturally determined, (the tribe's "fear" and ours) are not identical in respect of the attitudes which give rise to them. ...member of these societies, in prescribing "fear", are endorsing a response which, in so far as it is related to cultural beliefs and moral values, is distinct from, and not strictly derivable from, natural "fear". (this) does call into question the naturalists' account of socialised "fear" as explainable in terms of natural "fear" in virtue of their sharing the same qualitative features and causal conditions." (1986b p66)

While not exactly transparent, this passage gives important clues to what's going on in Harre and Armon-Jones' work. They seem to draw very different conclusions from the evidence of cultural variation to those I would be inclined to draw. I think the evidence shows that people are afraid of different things in different cultures. These differences reflect differences in beliefs and learning history. I am not alone in thinking this way. I can't actually think of a reputable modern emotion theory which wants to maintain that emotions are caused by the same things in all cultures, or whose theory rests heavily on the idea of normal or natural causes. Certainly neither cognitivism or affect-program theory does. Both give broad, functional definitions of appropriate elicitors - dangers should elicit fear, for example - and those definitions get filled in in the light of local conditions. Harre and Armon-Jones, however, take the evidence to show that different cultures have different states, which we wrongly translate as "fear". They want to taxonomise emotions in such a way that what causes them is relevant to their identity. If a state is caused by spiders and loose electric wires in one culture, but by scorpions and witchdoctors in another, then its a different state. Presumably, they would reject the theories I've just mentioned, saying that the taxonomy of emotions which abstracts away from the local significance of particular stimuli is inadequate. I have tried hard to find a justification for thinking this.
One such justification may lie in Harre and Armon-Jones concern to deny that there is any "inner essence" to emotions. There is nothing beyond the attitudes which cause them and the behaviour they engender. If this was true, it might be possible to mount an attack on the idea that there can be a constant emotion even when the causes of the emotion differ. One way in which Armon-Jones tries to show that there are no "inner essences" is by pretending that "inner essences" identifiable across cultures would have to be feelings, and then attacking the feeling theory of emotion identity with neo-Wittgensteinian arguments:

"...emotion feeling is constituted by those attitudes appropriate to the emotion. ...."fear feeling" would not remain unchanged (across cultures), but rather would be qualitatively different to the extent that the attitudes constitutive of the emotion feeling are specifically cultural" (1986b p66)

But the feeling theory is a straw man. The main candidates for the "essences" of emotions which remain constant across cultures are neural affect-programs and sets of beliefs and desires. Both of these will do the job of giving identity across cultures to emotions which are triggered in different ways, and even have wildly different effects. Armon-Jones offers no opposition to the affect-program view, but I imagine her reply would be that it can only deal with a small and unimportant class of emotional states, and that even in these cases the biological output is massively swamped by cultural overlay. Her reason for rejecting the cognitivist alternative is less readily discernible, but I suspect it may lie in her view of propositional attitude identity.

In her 1986b article, when discussing her view that feelings are sets of propositional attitudes, she says:

"..to "feel afraid" in (different) societies may involve to some extent, and on some occasions, natural beliefs. e.g., that x is menacing, and natural desires, e.g., to avoid x." (1986b p66)
But this does not mean that feelings are the same across cultures, she says, since on other occasions, fear occurs without these "natural" beliefs and desires. This would appear to be a contradiction of the essential claim of cognitivism, the claim which goes back to Kenny and beyond, that certain thoughts are essential to certain emotions. You can't, for example, be afraid of something unless you think it's dangerous, and want to avoid it. Such a denial might explain why she ignores the possibility that emotions might involve the same propositional attitudes across cultures, even if the particular causes of those propositional attitudes differ. But elsewhere, Armon-Jones states that her view is consistent with cognitivism:

"...constructionism would appear to require a theory of mind in which emotions, as instances of psychological states, are cognition based. ... This view that emotions depend upon appraisal (of situations as culturally appropriate to the emotion) is compatible with those contemporary philosophical theories that characterise emotions as cognitive in virtue of their dependence upon judgement and belief". (1986a p36)

It seems that Armon-Jones likes the idea that emotions involve thoughts, but is afraid that the analyses offered by the standard cognitivists (her "natural beliefs" and "natural desires") will make those thoughts culturally specific. But if this is what she thinks it's just off the planet! She would appear to believe that "thinking x dangerous" is an alternative to "thinking x has the magic power and malevolence characteristic of a Malay and is in a position to bring sickness and ill-fortune to you". In fact, the cognitivist means to give a broad ascription, of which the later ascription might be an instance. A conventional cognitivist will admit that Armon-Jones' different tribes have different specific thoughts when they're afraid, but claims that they all think something dangerous is around and want to avoid it.

1 See my discussion of the relation of cognitive and folk taxonomies in chapter two.
Chapter four: The Constructionist Theory

I am reluctant to charge Armon-Jones with such a simple misconstrual. The cognitivist is accused of wanting to analyse emotions with "natural" propositional attitudes, so perhaps when we see what "natural" means we'll see a better way to read her remarks on cognitivism.

In her 1985 paper Armon-Jones suggests that propositional attitudes may themselves be social constructions (p20). Some hint as to what she means by this can be gleaned from her 1986a:

"...emotions are constituted by non-natural attitudes, these being acquired in, and explicable by reference to, specifically sociocultural contexts ...such attitudes and their external referents are either irreducibly, or significantly sociocultural in nature." (1986a p36-7)

So natural and non-natural propositional attitudes are those whose objects are natural or sociocultural respectively. What does this mean? Unfortunately there are no examples at this point to clarify what's being said. Does Armon-Jones believe that there is a pre-cultural, pre-linguistic way of dividing reality, perhaps the way Romulus and Remus would have classified the world at the wolf's breast? On this view, beliefs about objects in the natural ontology would be natural. Or perhaps it's beliefs about the real kinds or universals on some theory of kind terms that are supposed to be natural? There is not enough evidence in the text to settle what Armon-Jones has in mind. Fortunately, we don't need to settle what she means to see that her natural/non-natural attitude distinction won't help her defeat the cognitivist version of the "inner essence" thesis. If natural and non-natural beliefs and desires are simply beliefs and desires about different objects or categories, then whatever account is given of those kinds and categories, it will still be a mistake to read the cognitivist as saying that fear is a set of natural beliefs and desires. When the cognitivists say fear is the belief that there is a danger and the desire to avoid it, they are giving a schema which can be fitted by many particular beliefs and desires,
including ones whose objects are as "irreducibly sociocultural" as it amuses you to make them.

The discussion of Averill and Harre/Armon-Jones should have shown just how diverse and inchoate a theory constructionism is at present. In Averill's case, the notion of a constructed response was entangled with a separate and contentious thesis about "improvisation". In the later pair of writers, the idea that emotion elicitors vary across cultures is confounded with the notion that emotions themselves vary. There are reasons why they think they can do without this basic distinction, but those reasons are far from adequate. In order to discuss constructionism constructively we are going to have to spell out our own version, which should be as stripped down as we can make it.

What is Constructionism?

Constructionism rests on two main observations. First, peoples' emotional behaviour, the way emotions are taxonomised, and the prominence of particular emotions and groups of emotions, seems to vary across cultures. Secondly, this variation is not random. The way in which emotions vary across cultures reflects differences in the way in which people live their lives in that society, and the way in which the society as a whole functions. The constructionist infers from the first of these observations that emotional behaviour is, to some extent, learnt. She infers from the second that emotions serve the purposes of individuals and societies. The results of these two inferences are put together in the following theory of emotions. Some (or all) emotions are learnt patterns of behaviour, which reflect the ideas and practices with which a person has been brought up. Because these responses reflect the society in which they were developed, there is a reciprocity between social practice and individual emotion, and the two are adapted to one another. The idea that emotions are learnt responses which serve our ends naturally leads to the suspicion that they are not involuntary behaviours.
Chapter four: The Constructionist Theory

The formulation I have just offered leaves a couple of important questions open. It offers no precise account of the relationship between responding emotionally and acting voluntarily. We still have to ask how having a constructed emotion relates to pretending, self-deceit and the unconscious. Another way to pose this question is to ask when an emotion is sincere on the constructionist account. My formulation is also uncommitted about the scope of constructionism. We still have to determine how much of the domain of emotional phenomena the constructionist account can be applied to. In the next section I shall discuss the first of these problems, the problem of sincerity. In the following two sections I shall discuss two problems for the view that all or most emotions are constructed.

4.2 The Problem of Sincerity

An emotional response that is constructed and implemented to fulfill social expectations sounds suspiciously like an insincere emotion. The constructionist has to give an account of why his theory is not just the theory that everyone is pretending. I expect the majority of people would agree with me that the pretence theory is one that it would, at least, be pleasant to be able to avoid. In this section I will discuss some constructionist accounts of insincerity. The first group are extracted from the work of Averill. The second group of theories, those of J.P Sartre and C.T Warner, suppose that all emotions are indeed insincere, but try to avoid the pretence theory by claiming that that they all involve self-deceit. They also hope that the notion of self-deceit will allow them to avoid talk of unconscious processes, which they regard as undesirable. Finally, I shall suggest that a conception of unconsciousness which is relatively uncontentiousness, and independently demanded by theories in many other areas of psychology will give us an effective solution to the problem of sincerity.
Chapter four: The Constructionist Theory

Averill's View

There is some evidence in Averill 1980b for an account of sincerity in terms of "involvement". Involvement is a theoretical variable. Emotions occur at points on a continuous scale of involvement. To some extent involvement corresponds to the pre-theoretical idea of differing intensities of emotion. Averill describes three general levels of involvement by way of illustration. First, he describes low involvement, where:

"enactment of an emotional role is largely a formality. An example is...the teacher who scolds a child for being naughty while inwardly laughing at its antics. The most common expression of emotion at this level is a simple verbal statement..... Such statements are not necessarily insincere just because they lack a great deal of personal involvement. They are generally appropriate to the situation and sufficient to achieve the desired end. In fact, if a person becomes too involved in an emotional role when the situation does not call for it, he is liable to be considered insincere or "affected" (Averill 1980B p316)

At a medium level of involvement physiological arousal becomes apparent. Averill claims that this is important because it symbolises a high degree of involvement to the individual. He claims that people frequently experience physiological arousal because they work themselves up in order to be "convincing" both to themselves and to others.

At a high level of involvement:

"the person may become so engrossed in the emotional role that he no longer seems in control of his own behaviour.... After the episode the person may claim that he did not know what he was doing or that he was not himself. When evaluating responses at this level it is important to


remember that it is part of the meaning of emotion to be "overcome", "gripped, "sized", etc. This meaning helps determine not only the kind of behaviour that will be exhibited, but also how the response will be experienced."(Averill 1980B, p317)

Averill's remarks about low involvement suggest to me that he would like to give an account of sincerity, or perhaps of some surrogate for sincerity, in the following way. An emotion is sincere when the response is appropriate to the elicitor. This would tie in with the basic constructionist theory, since it would make emotion displays sincere exactly when there is an emotion present. To have an emotion on the constructionist account is have the correct, socially prescribed response. It also picks up on a real strand in our ordinary thought about insincerity. We do feel suspicious of displays that are disproportionate or histrionic. But it will not do as an account of sincerity. It just fails to map onto our intuitive judgements of insincerity. Nearly all the "emotions" described in the above quotations seem intuitively insincere. The teacher is clearly pretending, and the social appropriateness of the pretence is just irrelevant.

Later in the same paper, Averill seems to realise that his account is implausible as a story about what we're actually conscious of. There are two suggestions in the paper as to how he might avoid this implausibility. The first is to make constructionism an account of the social evolution of behaviour patterns, rather than of any mental processes. The second uses the idea of the interpretation of behaviour as passive, which we have already alluded to. Averill expresses the first idea in the following passage:

"it is true that most people do not become emotional in order to fulfill some social obligation. But a role analysis is no more objectionable in this respect than is an analysis in terms of biologically based adaptive patterns. .... any specific episode of anger, love,...etc., may meet no social need. But if on
the average, or over the long run, such emotional syndromes conform to social norms, then their net result will be functional within the social system." (Averill 1980B, p336)

The way in which Averill treats the objection that people don't emote to meet obligations can be generalised to deal with the more general objections to talk of "acting out" emotional roles. Averill could say that this talk of "acting out" refers not to any mental activity on the part of the person with the emotion, but only to the function of the emotion in the relationship between the person and his society. Similar replies could be given to many other objections. But none of these replies bear much relation to what Averill was trying to do when he set up the position that is now under attack. They make the mentalistic way in which he originally laid out the theory unintelligible. To accept what Averill says in the above quotation and to use this strategy to meet the other criticisms of Averill's account would amount to abandoning the the entire theory. We would get a theory of how emotions come to have a functional role in a social system, but one which would say nothing whatever about the nature of the emotions themselves.

Averill's second strategy is to say that we interpret an emotional response as a passive one, rather than an action, because this is part of the role we take on when we become emotional. This seems to me a perfectly coherent notion, and ties in with the much publicised evidence for "confabulation" in psychological experiments. People can be induced to offer you their own theory of what went on in their mind, rather than an introspective report, while remaining convinced that it is the result of straightforward introspection. But interpretation or confabulation presume that what really went on isn't immediately available to the subject. If we're going to use such tools to render constructionism plausible, we need some account of how the real causes of emotional responses are hidden. The natural way to do this is to introduce unconscious processes.
Chapter four: The Constructionist Theory

For various historical reasons, constructionist theorists have tended to be hostile to the idea of unconscious states. Certain writers, notably Sartre, whose "phenomenological" theory is proto-constructionist, have attempted to bridge the gap between the prima facie reports of those experiencing emotions and the constructionist story about what's really going on, with the idea of self-deceit. Their solution to the problem of sincerity is to argue that all emotions are insincere. Because these accounts have had considerable currency among constructionist theorists, it is important to show that they don't offer a viable alternative to introducing unconscious processes, before offering our own account, which does introduce them.

Self-Deceit Theories

Sartre describes emotion as a form of magical consciousness. In becoming emotional we return to a way of confronting the world which we adopted as helpless children. Instead of confronting the world as it is and taking appropriate action to deal with it, we distort or annihilate our perception of it. Thus I faint in terror so as to remove the object of terror from my consciousness. The whole essence of emotion is self-deceit. We are emotional about what we cannot face up to, and being "emotional" consists precisely in not facing up to it. The theory encounters an obvious difficulty with pleasant emotions, such as joy. Sartre replies that we must see joy as an attempt to possess the object of our delight completely and instantaneously, to drain all the pleasure from it at once, rather than taking the trouble to slowly and prudently enjoy it. Thus, in ecstatic love we eschew the infinite care which is needed to make a relationship succeed in favour of an immediate rapture. The theory also has trouble with mild emotions. Sartre denies that these exist. Emotional experience, he asserts, is always of absolutes. In emotion the object's emotional quality becomes its central feature for us, and it is

1 Since Sartre has declared love to be an impossible enterprise full of internal contradictions (Sartre "Being & Nothingness" pp361-413) this sounds like a rational and sensible approach, but he doesn't make the connection.
apprehended only under its emotive description. He accounts for our experience of weak emotions as glimpses of strong ones at a distant remove from us. I find this unconvincing, to say the least.

Sartre has a long-standing dislike of unconscious states. He sees positing an unconscious as just one more way to escape from the conclusion that we live our lives in a state of self-deceit and "bad faith". So we must presume that the various states referred to in his account of emotion are present to consciousness. Unfortunately, they aren't. Any credibility Sartre's account may have would come from focusing our attention on situations in which we are "working ourselves up", as when we consciously use anger in a marital quarrel or on a committee, to get our way. I don't deny that such cases exist, but I do deny that when my parachute fails, my fear is a state I put on in order to avoid having to face up to my problems. To prove this point I can only appeal to the reader to consider how much of his or her own emotional life he/she can seriously apply a Sartrean account to. If Sartre thinks that the notion of self-deceit will fill this hole in the account he's just mistaken. If he wants his "self-deceit" to be something we are, or could be, conscious of, then his account isn't experientially credible. The alternative is for self-deceit to involve the use of unconscious states and processes, and he rules that out.

I now want to discuss C.T Warner, who is well aware of the problems we've just canvassed, and attempts to offer a substantial account of self-deceit that avoids mentioning unconscious states or processes. Warner (1986) offers an account of anger as an essentially self-deceptive condition. In his own words,

"Fundamentally, (anger) is her act of self-misconstruing that she is misconstruing."

This needs quite a bit of unpacking. My own, rough, understanding is as follows. An angry person is reacting to a supposed wrong. It is of the essence of anger to be disproportionate and exaggerated. The prime drive of an angry person is, therefore,
to deceive themselves about how unreasonable they are being. Their anger is self-sustaining, because they interpret any attempt at conciliation as further hostility (Warner calls this "resistance"). The exposition is complicated by Warner's tendency to discuss reciprocal anger throughout, where both parties are enraged at the exaggerated and unreasonable anger of the other. It is supposed to be a problem for all previous accounts that they assume that anger is a response to a substantive injury, whereas Warner seems to maintain that it is the absurdity of the first person's anger that provokes the other's anger and so on ad infinitum.

Now, as a general theory of anger, this doesn't even get off the ground. Warner has modelled his whole account on the marital quarrels of modern Americans, as his examples show. They all feature reciprocal marital anger where at least one party is being unreasonable. Warner claims that examples of similar "accusing and self-victimising" emotions could be found in many cultures. But even if this is true, it has no bearing on the many instances of anger in our own culture which aren't part of a marital power struggle. The idea that all anger is a matter of exaggeration and self-victimisation falls apart the moment you take it outside this very limited arena. If Warner thinks that a Russian partisan who's relatives are raped and murdered by the Nazis responds by victimising himself, or exaggerating, then he's got a warped mind. He does admit at one point (p143) that there may be anger which isn't self-deluding, and gives Christ's anger at the money changers in the temple as an example. I doubt whether this constitutes hard data, but I'm quite happy to grant him the general point. Once Warner has made this admission, the real question is how much everyday anger will fit his model? Warner gives us little to go on in answering this question, since his arguments for the theory are almost entirely a priori, and the only empirical evidence in play is, presumably, Warner's own experience of marital quarrels. My own opinion, for what it's worth, is that its application will be limited, and certainly too small to justify the title of a "theory of anger".
As we said at the outset, Warner's theory is accompanied by an account of self-deception. This is important, because even if we restrict Warner's story to cases where it is plausible to suppose that people are putting their anger on, or exaggerating it, we still need an account of how they can fail to realise this, or our account will be restricted to the still smaller group of cases where people are explicitly pretending. Like Sartre, Warner is extremely hostile to unconscious processes. He tells us that they are self-contradictory, and involve conceptual confusions, although this is not spelt out in detail. Because of his opposition to the unconscious he wants to avoid the idea that self-deceit is a matter of the relative accessibility of one belief compared to another. The angry person doesn't have inconsistent beliefs, they just have a weird style of belief formation.

"On the agentive view, her self-deception is not a matter of concealing a belief, but a matter, we might say, of believing perversely." (1986 p164)

One thing which is very clear in Warner's text is that this view is supposed to stand in sharp contrast to views which use the unconscious in their analysis of self-deceit. I have tried to find the supposed contrast, and, apart from vague references to the incoherence of accounts which use the unconscious, can only find the argument embodied in the following passage:

"...the problem with the classical conception of self-deceit is the assumption...(that) what the self-deceiver resists (e.g., the possibility of her own malice and fraudulence) is the very belief about which she deceives herself. On this assumption, the only mode of self-deception available is concealment from herself - e.g., relegation of a belief content to a "level" below consciousness. .... On the agentive view, on the other hand, what she resists in her anger...is not what she is deceiving herself about. It is not what she would believe upon coming out of self-deception. It is an artefact
of the anger itself, and it would not be a possibility - it would disappear - without the anger." (1986 p164)

So what is it that she deceives herself about? Warner drops a few dark hints. It is something "not possible to deny (or affirm) apart from the self-deception". It is, as quoted above, "believing perversely" which is itself explained as:

"...a matter of participating insincerely (sic) in that form of life in terms of which she learned to maintain herself as a person"

I find these remarks totally incomprehensible, so I venture to offer my own version, which has the advantage of being coherent, and seems to fit reasonably well with some of the text. I read "believing perversely" as having weird dispositions to form beliefs. We can then take what she is "resisting" in the longer passage above to be the suggestion of her "malice and fraudulence", while what she deceives herself about is the fact that she has been grievously wronged by the object of her anger. This reading has the advantage of making the object of self-deceit the same thing "she" was supposed to be deceiving herself about in the discussion of anger earlier in the paper.

The only problem with this reading is that it makes it hard to see why so much turns on the idea that what she deceives herself about isn't what she is "resisting". We can have a "classical conception" of self-deceit, and keep this feature. Suppose she believes she's been wronged when she has good evidence that she hasn't, and the motivation for this is a set of beliefs and desires existing at some level inaccessible to consciousness. Then what she is deceiving herself about is her being wronged, but what she is subconsciously resisting may be suggestion that she is malicious and fraudulent. I am therefore unable to see why we cannot use the unconscious in our analysis of self-deceit, and will proceed to do so.
Sincerity with an Unconscious

I now want to outline an account of sincerity which utilises a relatively uncontroversial conception of unconsciousness. I hope this account will get us everything we need for constructionism, without running us into the kind of absurdities we've just encountered. My proposed account of the difference between conscious and unconscious states will lean heavily on the computer analogy. It is in line with views of the unconscious expressed by, among others, Daniel Dennett (1978, 1982). Basically, the proposal is that all mental processes proceed on the same level, but some of them are monitored by second-order states. Thus, as Jerry Fodor has observed, various stages of our perceptual input systems are wholly inaccessible to consciousness, whereas the output of those systems is not (Fodor 1983). Similarly, the analysis of an object that leads to an immediate fear or anger response is often unconscious, whereas the processes that lead to the slow realisation that something is dangerous or annoying are often accessible to consciousness.

One of the advantages of this model for our purposes is that it does not make the distinction between conscious and unconscious processes too rigid. A process capable of being consciously monitored may, for example, proceed unconsciously because of an unconscious, or even a conscious, decision not to monitor it. This happens when we direct our attention, or have it compelled by some striking feature of the environment. Other mental activity, such as reception of information by the other sense modalities is left unmonitored, although it may still be playing a vital role in such important activities as navigating our way around the room. The same kind of thing may happen with the monitoring of processes leading to emotional responses.
With this concept of the unconscious at our disposal we can give an account of the sincerity of c-emotions. A c-emotion is sincere simply when the subject is unaware that it is a c-emotion. That is to say, he or she is unaware that they are only exhibiting the behaviour in order to fulfill the stereotype for that emotion. They are not aware of working themselves up into a state of physiological arousal, or of pretending that they cannot help the state they are in. This general requirement may be met in several different ways, which is appropriate to the concept we are trying to explicate. Our account will allow emotions to be more or less sincere, depending on the extent to which the subject realises that his responses are voluntarily initiated or exaggerated. It also allows us to distinguish straight pretending, self-deceit, mere inattention to ones motivation, and real deep-seated inability to get at ones motivation. At least this many cases of emotional insincerity are recognisable and it is a virtue of an account to be able to fit them all in. In straight pretending all or most of the processes leading to the c-emotion are monitored by consciousness. Self-deceit probably covers a number of cases, among them are cases in which the subject would be conscious of the processes leading to the c-emotion, if they had not been consciously or subconsciously motivated to direct their awareness elsewhere. It may also be possible for a self-deceiving subject to hold contradictory beliefs about the c-emotion. In the case of inattention the subjects unawareness of the processes is due to the directing of their attention elsewhere and such things as another person pointing their motive out to them or a moments reflection will be enough to make them give up the emotion or change to pretence or self-deceit. A further important class of cases, which overlaps with the category of self-deceit, is deep-seated unconsciousness, where the states which lead the subject not to monitor the processes are themselves not easily available to consciousness. This is the kind of case in which the emotional response really is highly analogous to the symptoms of an hysterical illness. There is a still deeper case of unconsciousness, in which it results from broad architectural features of our information processing.
arrangements. But there is no evidence I know of which suggests that the processes leading to c-emotions can be unconscious in this way\(^1\).

4.3 The Problem of Inappropriate/Sinful Emotions

It seems that we can solve the problem of sincerity by abandoning the unnecessary constructionist hostility to unconscious mental processes. In this and the next section I consider some more difficult problems. I shall show that constructionism must be restricted to a sub-class of emotions if it is to avoid these.

One reason to believe that not all emotions can be constructed is that people have emotions which are hard for other members of a society to understand, and emotions which that society condemns. People laugh at funerals, and fear harmless furry animals. These are the inappropriate emotions. Other people are easily moved to hatred, envy or jealousy. These are the sinful emotions, easily understood, but generally condemned. There can be emotions which are both inappropriate and sinful, such as the joys of the sadist. Sinful and inappropriate emotions are prima facie problems for the view that all emotions are learnt patterns of behaviour which reflect, embody and reinforce the values and structures of the society which shapes them. Armon-Jones has devoted considerable space to an attempted constructionist explanation of these two phenomena. I shall try to show that the problem of sinful emotions is largely of her own making, and is easily avoided. I shall then try to show that the arguments she puts forward in her account of inappropriate emotions collapse her constructionism into a modified cognitivism, and make nonsense of her radical manifestos about the cultural construction of emotions and their lack of any cross-culturally definable inner essence. I conclude from this that the attempt to stretch constructionism to cover these problem cases was a mistake.

\(^1\) Processes leading to affect-program responses may be unconscious in just this way. See section 3.5.
The Sinful Emotions

There is a fairly simple reply to the sinful emotions problem that runs as follows: when we say that an emotion is prescribed by society for certain situations, we use "prescribe" to mean "regard as appropriate for a normal person" not "regard as the right thing to do". Clearly something can be "prescribed" in the first sense and not in the second. We can admire Gandhi's loving his enemies while still regarding this as deviant behaviour. The very way we talk about it will teach our children that it is not expected of ordinary people. But Armon-Jones doesn't use this defence. In fact, it appears from her 1986b (p71-6), that she doesn't accept the sort of distinction I've just drawn. She appears to think that when people learn which emotions to display, they learn the morally prescribed emotions for that situation. To maintain this position she argues that the sinful emotions are actually prescribed in a moral sense by society. In some cases they are directly prescribed: a good Republican should hate communists. In other cases, they are necessary concomitants of other responses that are directly prescribed. She argues that someone who was not jealous of their spouse would cease to value them in the required way, and envy, she says, is only an extension of admiration and the desire to emulate achievement. Someone who has these prescribed emotions to the correct degree will have envy as an inevitable consequence. The underlying thought here seems to be a cognitivist one. The emotions involve propositional attitudes, and the logical relations between those attitudes mean that having certain emotions requires having others, which may be undesirable in themselves.

These suggestions of Armon-Jones' would deal with some sinful emotions, but there's no reason to believe that they'd deal with all, or even most. It may well be the case that Republicans should hate communists, but they frequently hate fellow Republicans too, and that is neither prescribed, nor, so far as I can see, a necessary consequence of what is. Perhaps Armon-Jones believes that an emotion can be a constructed response, as long as some special subset of its instances occur in
morally prescribed contexts. This is suggested by other arguments in the same section. She argues that some of the sinful emotions are condemned only when carried to excess, and have moderate instances which are morally prescribed. She points out that in a different cultural context, even the extreme case of the sinful emotion may be morally prescribed. Unacceptable jealousy in an Australian might have been quite normal in a medieval Arab. All this seems to aim at showing that even for the most sinful emotion, there are some instances, somewhere, that are morally prescribed. But I just don't see what Armon-Jones thinks this will prove. She has asserted that an emotion consists in some set of culturally prescribed attitudes elicited by culturally appropriate construals of the stimuli. She has ignored the possibility of a non-moral use of "prescribed", and I just can't see how an emotion which has even one instance that is universally condemned can fail to be a counterexample. However, since the simple defence I outlined at the outset is still available, I don't think sinful emotions have to be a problem for constructionism in general.

The Inappropriate Emotions

I now want to turn to Armon-Jones' treatment of the inappropriate emotions. She divides inappropriate emotions into two kinds (1986b p71). In the first kind, the emotion is perfectly understandable, given the beliefs of the agent, which happen not to coincide with ours. They believe that small furry animals live only to bite, or that in a wide open space the sky may fall on you, or that children ought to have sexual relations with adults. These emotions are no problem for the constructionist, as long as they can find the right beliefs. Where they can't, they adopt a version of the strategy adopted by many cognitivists to deal with apparently objectless emotions: in the face of an apparently inexplicable response, posit beliefs which rationalise it. The sole evidence for these beliefs is the existence of the response, and the need to explain it. The main objection to this strategy is that in some cases beliefs so ascribed don't fit well with the subjects other beliefs, and the subject may
Chapter four: The Constructionist Theory

repeatedly refute the ascription in thought and deed. An agrophobe may deny believing that the open space is dangerous, and happily send their beloved children into the space, and still become petrified with fear if they have to go out into it themselves. When this problem arises, Armon-Jones has an alternative account of inappropriate emotions to put forward. In Armon-Jones' second case, the inappropriate response is explained as a leftover response which was culturally prescribed to childhood, or the result of a deviant upbringing which may have inculcated emotions that were not culturally prescribed at any stage. In such cases, she says:

"The second case in which an emotion involves culturally inappropriate attitudes can be explained by the constructionist as non-paradigmatic, and as liable to deprive the agent of his status as someone in possession of normal sensibilities." (1980b p71)

Armon-Jones is treading on very thin ice here. She has been trying to show that emotions are socially constituted, and, that their identity depends upon the set of cultural standards that constitute them. But if this was the case, then people who lacked the right social training would not have the same emotions in inappropriate contexts, they would have no emotions at all, or at best, strange emotions, which their compatriots would be at a loss to name. Yet it is fairly clear that Armon-Jones does want to say that the same emotions can exist in a deviant contexts. She shows this on page 71, when she says that we judge people to be abnormal if they have emotions of inappropriate intensity, or which involve incomprehensible attitudes. From a theoretically uncommitted viewpoint this seems to be the right thing to say.

Members of the Paedophile Information Exchange find sexual relations between adults and children delightful, and pleasant to contemplate. What is strange about them is that they feel these emotions rather then those we would feel in their shoes. We do not contemplate them like Mr Spock observing the earthmen, unable to comprehend what these things they call "emotions" are! But although this seems to
be the right account to give of the inappropriate emotions, by giving it Armon-Jones is saying something radically at odds with the rest of her theory. She is admitting that an emotion can retain its identity even if it occurs without the cultural attitudes which she elsewhere takes to constitute the emotion. In this passage, her theory of emotion seems to be a form of cognitivism.

If Armon-Jones were to consistently extend the approach she takes in this passage, her theory would simply combine the cognitivist analysis of emotions as sets of propositional attitudes with some uncontroversial but much neglected observations about cultural and individual variation. Which experiences cause, and what action results from, certain propositional attitudes varies hugely with local conditions, as a result of other locally prevalent propositional attitudes. The causal relations of propositional attitude systems are, after all, holistic. Fear might be almost unrecognisable in another society unless you had a comprehensive grasp of local conditions and the culture of the inhabitants, even if fear was the same sort of propositional attitude set in all cases. This is an interesting and important variation on cognitivism. But surely it is not what Harre and Armon-Jones want? Their talk of constructing emotions, of incommensurability of emotions across cultures, and the cause-involving taxonomy discussed above suggest something far more radical.

In summary, Armon-Jones' response to the problem of inappropriate emotions has been to allow emotions to have purely cognitive identity conditions, in order to get identities between standard and deviant cases of particular emotions. I believe that the proper response would have been to admit that constructionism gives only a partial account of emotion phenomena, and concentrate on a detailed elucidation of what's happening in cases where it is not plausible to suppose that what we have is a readily understandable cognitive core, altered in its outward manifestations in response to local conditions and beliefs. In my own reconstruction of constructionism, I shall be primarily interested in a constructionism that can deal with phenomena beyond the scope of standard cognitivist theories.
4.4 Problems of Universality, Innateness & Passivity.

The single most important reason to deny that all emotions are culturally specific learnt behaviours is the large body of sound empirical evidence showing that some of them aren’t. We have already reviewed this evidence at some length in chapter three. For our present purposes it can be summarised under three heads, universality, innateness and passivity.

Universality

First, there is the evidence of emotion universality. Well defined patterns of facial behaviour occur in vastly different, and even visually isolated, human societies. These behaviours occur in response to functionally equivalent stimuli in the various different settings. The fear behaviour occurs in response to dangers, the anger behaviour in response to challenges, sadness in response to loss, however danger, challenge and loss are instantiated by a given environment and body of beliefs or experiences¹. These results suggest very strongly that there is a common core to the emotions of fear, anger, etc, in all human societies.

Innateness

The evidence of universality gives some support to an innateness hypothesis. But there is also more direct evidence for innateness. Eibl-Eibesfeldt has shown that the universal facial displays are present in children born blind, and thus incapable of learning them². Trevarthen (1984) reports sensitivity to facial displays of emotion in 36 hour neonates. This suggests that the infants sensory apparatus, in many other ways underdeveloped, is primed to receive these vitally important signals

¹ See section 3.2.
² See section 3.2
from its caregiver. Thus, both the display and recognition of the basic, emotionally significant facial responses appears to be innate.

Passivity

Other evidence suggests that the innate displays are resistant to learning, and are overlaid, rather than removed, by later cultural training. The displays can be shown to occur even in the famously "inscrutable" Japanese, by the use of secret observation, and slow motion analysis of instantaneous responses. The effect of cultural training appears to be to create habitual learnt responses which serve to block out the innate response before it can normally be observed. This phenomenon of overlaying suggests that the systems which mediate the innate response are not the same as those which mediate voluntary facial movement. This would count strongly against the constructionist view of emotional behaviour as agentive, and suggest that traditional theories were correct to view at least part of our emotional response as passive behaviour governed by processes other than our ordinary system of belief(desire action rationalisation. In chapter three I suggested that we can restate the traditional idea of passivity using a Fodor-style notion of modularity.

The standard constructionist response to this kind of attack is to deny that the responses studied in the experiments are "really" emotions. We saw an example of this in section 4.1, where Averill attempted to exclude what he called "natural fears". We also saw that this is merely stipulative. It is, of course, open to the constructionist to make this stipulation, but to do so is to admit the substance of my claim that constructionism cannot give a total account of the emotions.

---

1 See section 3.2 and Appendix I.
4.5 The Real Role of Constructionism.

Sections 4.3 and 4.4 have shown that constructionism cannot hope to be a total theory of emotion. We saw that the attempt to extend it onto the inappropriate emotions led to an abandonment of the most radical and interesting theses of constructionism in favour of a modified cognitivism. We have just seen that the affect-program phenomena are wholly at odds with any constructionist account. In this section I want to show that the real use for constructionism is to give an account of just those phenomena which escape a standard cognitivism treatment. I shall consider three kinds of phenomena, constructed overlays of other emotional responses, constructed tokens of emotion types not all of whose tokens are constructed, and constructed emotion types.

**Constructed Overlays**

The first of these, constructed overlays, is the least controversial. Suppose that there exists a response caused either by an affect-program or by a complex of propositional attitudes. In each case, the response will have behavioural manifestations, in the one case those hard-wired into the hypothalamus, and in the other case, those rationalised by the attitudes, in conjunction with the rest of the persons beliefs and desires. Let us call this the "basic" behaviour for that response. It seems simple enough to suppose that culturally specific responses could become closely associated with the basic behaviours. In the affect-program case, it is indeed simple enough. As Ekman’s experiment on the Japanese students showed, learnt "display rules" about when to hide emotion, and substitute behaviour such as a polite smile, can become as automatic as the affect-program itself, and can, with sufficient enculturation, make the affect-program response virtually undetectable to normal observers. There seems no reason to suppose that behaviours enhancing, or adding to, the basic behaviour could not be equally automated.
With cognitive emotions things are a little more difficult. The problem arises when we try to distinguish the constructed element from behaviours rationalised by cognitive core of the emotion, in combination with locally or culturally specific attitudes. Anything which is so rationalised is, of course, part of the basic behaviour. Consider as an example, the explanation of the alleged behavioural patterns of injured or insulted Koreans. Allegedly, Koreans become increasingly respectful and deferential when they are wronged. A radical constructionist might want to push the line that they have a quite different emotional life, in which injury and disrespect have a different significance. Actual constructionists (See Warner 1986 p165) prefer to say that in Korean society, this is the most effective way to punish someone, since it implicitly asserts your moral superiority, and makes their behaviour seem more shameful. This seems more plausible, and it fits in nicely with a standard cognitivist analysis of anger. Anger is the belief that you have been wronged, and the desire for retribution. The Korean has some set of propositional attitudes that fits this schema, and suitable beliefs about his society to make it rational to use deference as a tool to revenge him or herself. Unfortunately, this story seems to make the culturally specific, Korean behaviour part of the basic behaviour associated with the emotion. It might count as constructed behaviour on the minimal version of constructionism put forward by Armon-Jones in section 4.3, but, as we saw there, that just ditches most of what's interesting about constructionism. I'm not suggesting that there's anything wrong with making the Korean's response part of the basic behaviour, but it does show us that finding things that aren't basic behaviour is going to be harder than you might think. What we need is a constructionist story about how learnt elements could come to operate in conjunction with, but not merely as part of, a cognitive response. In other words, something to parallel the way in which they might overlay an affect-program response.
I can think of two phenomena that might fit this need. The first is behaviour rationalised, not by the attitudes that constitute the (cognitive) emotion, but by the desire to live up to the full cultural paradigm of that emotion. Person P realises that she's been wronged, desires revenge, and so is angry. Although Korean-style deference might actually be more effective, she shouts and screams, because that's how she thinks an angry person ought to behave. On the story about constructionism given in 4.2, this will be a constructed response, and it will be sincere to the extent that its conformist motivation is unconscious.

The second case is where certain behaviours are habitually associated with others, although they are not rationalised by the subjects beliefs and desires. Thus, someone might habitually clench their fist when they raise their voice. Where such habits fit in with the cultural model of the emotion involved, there is likely to be some story about how they got established as habits which involves pressures to conform to cultural norms. The paradigm case may turn out to be one in which the behaviour is originally motivated by the need to conform to a cultural model, but then comes to be automatically produced in association with other anger behaviour. In conclusion, it seems that we can find behaviours which could reasonably be regarded as the constructed overlays for cognitive emotions. The behaviour in question would not be rationalised by the beliefs and desires that make up the cognitive emotion, without additional, perhaps unconscious, beliefs and desires which define a cultural paradigm of the emotion.

Admittedly, this sort of scenario is not a huge variation on standard cognitivism. Emotions are still constituted by cognitive states. But it is an important variation, because a sensitivity to these cultural factors will complicate our taxonomy of emotions. The standard cognitive taxonomy may not be of much use in a culture where its categories are cross-cut by local ways of classifying and thinking about mental states. Suppose there are two radically different cultural paradigms of how to behave when you think you have been wronged and desire
retribution, related to two sorts of circumstances and leading to two utterly distinct
sets of behaviour. On the cognitive taxonomy, both these states are anger. So the
taxonomy of emotions required to give an account of the role of the several
emotions in that society may not be the standard cognitive one.

**Constructed Token Emotions**

The second kind of phenomena that might call for a constructionist treatment is the
existence of whole token constructed emotions. In such a case there exists an
emotion type some of those tokens are constructed emotions. I shall call such
things token c-emotions. For an event to be a token c-emotion, we have to eliminate
the possibility that the behaviour is caused by an affect program, and the possibility
that what we are observing is a cognitive emotion. The first requirement is fairly
easily met, the second is more puzzling. We would need a situation where the
behaviours of the token c-emotion are caused by some set of propositional attitudes,
but where these propositional attitudes are not such as to allow an ordinary
cognitivist analysis of the emotion. The obvious case where this occurs is where
someone pretends to have an emotion. Justinian behaves as if he is angry with
Theodora, but the propositional attitudes rationalising this behaviour relate to the
need to bring John of Cappadocia's hostility to Theodora into the open so that he
can be destroyed. Such a case is unproblematic. But a c-emotion is, as we saw in
4.2, a more or less unconscious pretence of emotion. So a token c-emotion will
occur whenever someone is unconsciously motivated to put on an emotion, or
exaggerate their emotional state. Theodora rages against Belisarius, although he has
served the empire magnificently, and the perceptive historian sees that her goal is to
prevent him gaining too much popularity with the people by tainting him with self-
agrandisment and treason. However, he also sees that her corruption is so exquisite
that she does not need to consciously "put on" her anger, her emotional displays
follow her political needs without any conscious prompting. Less calamitous
instances of token c-emotions might be found in the anger of C.T Warner's modern American couples in section 4.2.

Once again, we are still analysing emotions with cognitive states, but they are not the cognitive states envisaged in standard cognitive theories. A token of c-anger is not an instance of anger because it shares the standard cognitive content, but because it is an attempt to mimic anger as it is recognised in that culture.

I have treated constructed overlay's and token c-emotions as two different phenomena, for ease of exposition, but the distinction is not sharp. Where the trigger for a display of c-anger is a mild but genuine resentment, it will be a delicate, and perhaps stipulative matter to distinguish constructed anger from cognitive anger with a constructed overlay. It seems likely that the sorts of mental processes we are now discussing are going to come in all sorts of wonderful varieties, all doubtless of infinite interest to the sensitive novelist. A general classification of these phenomena, however, especially in the language of folk psychology, is not a very promising research project.

**Constructed Emotion Types**

The third and final phenomena we have to consider is the full blown c-emotion, a constructed emotion type. Such a thing would be closely akin to the "wild pig" phenomenon of the Gururumba, described above. It would simply masquerade as an emotion, rather than a mental disorder. Some authors, notably Averill and Boothroyd (1977) have maintained that romantic love is such a phenomena. They suggest that the behaviour observed in full blown romantic love does not result from any set of propositional attitudes that constitute the attitude "loves". Rather, the behaviour is an attempt to conform to cultural or sub-cultural models of how to behave at a certain age, and in certain circumstances. The idea is an appealing one, but turns out to be very hard to provide conclusive evidence for.
Chapter four: The Constructionist Theory

There are four conditions that would have to be met in order for us to conclude that romantic love is a constructed emotion. Firstly, love must not turn out to be an affect-program phenomena. It seems likely enough that it won't, at least as long as we're talking about something distinct from immediate responses to the opposite sex. Secondly, it would have to turn out to be false that "A loves B" can be analysed as a set of propositional attitudes, or even as a family of sets of attitudes. Thirdly, romantic love must turn out to have the right kind of underlying structure. The motivation of those romantically in love must turn out to be a drive to conform to a cultural model. Finally, if it is to be a sincere emotion, rather than a pretence, this motivation must be less than fully conscious.

The problem comes when we try to meet the second condition. Attempts to provide cognitive analyses of love are notoriously unconvincing (Newton-Smith 1973 for instance), but their failure is consistent with love being a cognitive state with cultural overlay, or with token c-love being very common. I find the idea that love is a highly conventional phenomena pretty plausible, as many people have, but once again, this might be a matter of cultural overlay or frequency of token c-love. I suspect that this situation would be repeated for most putative c-emotion types. It will always be hard to eliminate the possibility that there is a core of propositional attitudes which constitute an emotion type of which other cases are token c-emotions. This problem is made worse by the fact that so much cultural variation can be accounted for within the cognitivist model. Even if an emotion type is confined to one culture, and seems inexplicable to outsiders, it may still be a cognitive emotion, even a common cognitive emotion, interacting with outlandish local beliefs and desires to give the apparently incomprehensible behavioural output.

---

1 La Rochefoucauld, in "Maxims" famously remarks "No-one would love who had not read of love", and this line of cynicism can be found as far back as Chaucer's satire of the medieval tradition of courtly love in the "Miller's Tale".
General Observations on C-Emotions

We have seen three kinds of phenomena which constructionism may be able to handle, and which are beyond the reach of both affect-program and standard cognitive theories. There are a couple of important facts common to all three cases. First, c-states are more directly linked to particular behaviours than standard cognitive emotions. A token c-emotion is anger, rather than fear, just because it mimics stereotypical anger behaviour in that society. In the case of a c-emotion type, if there are any, the culturally specified behavioural stereotypes associated with that emotion term are all there is to the emotion. To have the emotion just is to be acting out that learnt cultural model. The fact that behaviour is so essential to c-emotions may explain why the doctrine that all emotions are constructed is attractive to neo-Wittgensteinian philosophers like Armon-Jones. If she think that all mental states are criterially, or even constitutively, linked to behaviour, then constructionism may appear to be the application of such views to the emotions. Such a philosopher will also as inclined to think that constructionism will be straightforwardly compatible with cognitivism, as propositional attitudes are themselves destined to be analysed behaviourally. Some of what Armon-Jones says suggests that she thinks this. In this paper I have restricted myself to showing that if you're primarily interested in a theory of emotion, universal constructionism doesn't look like a good candidate. If you're primary interest is in developing a criterialist philosophy of mind, an approach which you support on independent grounds, different arguments would be needed.

A second common feature of the three uses of constructionism we've outlined is that the constructed phenomena mimic other kinds of emotional phenomena. The original constructionists, who don't think there are any other emotions, express this by saying that they are modelled upon each culture's picture of what emotions
Chapter four: The Constructionist Theory

should be, and vary with that picture as it changes across cultures (Heelas 1986). Where each culture's picture comes from is a question which ought to be asked, but isn't. The idea that c-emotions are imitations of emotions makes a good deal better sense once you admit there are some other emotions for them to imitate. This is not to deny that cultural stereotypes of emotion may vary, and that local c-emotional phenomena may vary with them. But it is to insist that the cultural stereotypes rest on something. I suggest that the features of cultural stereotypes of emotional behaviour have a lot to do with the real features of the affect-program emotions. In particular, the idea that we are passive subject of emotion may derive from the modular nature of the affect-programs.

The idea that emotions are passive may be central to an explanation of the c-emotion phenomenon. Averill argues that we construct c-emotions because emotions, as supposedly passive, unavoidable states, get special privileges for those who "suffer" them. This certainly seems plausible for many token c-emotions. Anger makes certain behaviour acceptable that would be condemned if executed in cold blood. It would also explain why people should go to the trouble of constructing whole c-emotion types. The Gururumba's constructed spiritual disorder helps individuals cast off social ties, and allows society to reduce the implicit challenge to its moral order by blaming the event on ghosts. Similarly, social ties can be cast off with the assertion that one is carried away by love.

4.6 Summary of Conclusions

In section one of this chapter, I tried to show that constructionism as it now stands is a diverse and inchoate group of doctrines, obscured in the works of its major exponents by unnecessary, and often controversial additions. The essence of constructionism lies in two important observations. First, emotional behaviour, and the way that people classify that behaviour, varies across cultures. Secondly, these
Chapter four: The Constructionist Theory

variations in emotions between cultures are intimately related to other facts about the functioning of those cultures. Constructionism is the attempt to explain these observations by the theory that emotions are, at least partially, learnt patterns of behaviours reflecting and reinforcing the values and structures of the society in which they occur. To make this doctrine more concrete, consider anger in western civilisation. A constructionist theory might argue that the people learn at least some aspects of their usual repertoire of anger behaviour from those around them, and that the stereotypes they acquire reflect local beliefs, such as that anger expresses itself in obscenities and violence, or that it is an involuntary response, and often can't be controlled. The theory might further argue that anger serves the individual by licensing displays of aggression which would otherwise be condemned, and serves society by allowing it to excuse behaviour which it is not inclined to punish, but does not wish to explicitly permit, such as violence towards adulterers.

In the main body of the chapter we tried to free this theory from the encumbrances foisted on it by the various major constructionist authors. In 4.1 we saw that Harre and Armon-Jones consistently refuse to distinguish the thesis that we learn how to behave when angry from the thesis that we learn what to be angry about. The later thesis is totally uncontroversial, and the "naturalist", the constructionist bogeyman who denies this thesis, is a straw man. This mistake makes much of the "evidence" offered by these authors irrelevant, since the learning of emotion elicitors is compatible with such paradigm anti-constructionist views as the affect-program theory! We saw that Armon-Jones has doctrines that are supposed to remove the need to distinguish the two learning theses, but we also saw that these doctrines don't do the job.

In section 4.2 we saw that many constructionists make their view unnecessarily implausible by denying that the construction of emotional responses is an unconscious process. By resisting this idea they collapse constructionism into the theory that everyone is pretending. Some theorists try to cover the gross
implausibility of this by calling on the idea of "self-deceit". Unfortunately, this denotes a phenomenon at least as problematic as that which it is supposed to explain, and all adequate theories of which seem to involve the unconscious processes these authors are so keen to avoid. We also saw that a relatively uncontroversial conception of unconsciousness will allow the constructionist to escape all these problems. Once we admit the unconscious into our account, we can make sense of the way in which constructed phenomena inhabit a space between intentional action and truly passive behaviours, such as reflexes or affect-program responses. The learnt behaviour displayed in a sincere c-emotion is not performed with the conscious intention of conforming to a cultural model. Instead, the subject's learnt conception of how to behave subconsciously motivates the behaviours that make up the response. These behaviours may also become automated, or co-automated with other behaviours, and here the sense in which they are "motivated", even unconsciously, is minimal.

In section 4.3 we saw the disastrous effects of trying to run a constructionist theory for all emotions. In her constructionist account of inappropriate emotions Armon-Jones collapses her view into a modified form of cognitivism. There is nothing wrong with this view in itself, but it is not at all what she promises us elsewhere. Furthermore, it is not a view which can deal with the cultural phenomena that elude a cognitivist analysis, and which we might hope constructionism would deal with. In section 4.4 we saw that constructionism can never hope to deal with the emotional phenomena discovered by research into the psychoevolutionary/affect-program theory. In my own sketch of constructionism I have presented it as a tool to deal with cultural variation in emotion, where this escapes any standard cognitivist analysis. My constructionism has no pretensions to encroach upon the proper domain of the affect-program or cognitivist theories. It is, however, a useful element in an overall account of emotional phenomena.
5.1 The Need for Multi-Vocality

In the preceding three chapters we have encountered an astounding diversity of emotion theories. In this chapter I want to argue that this diversity in theory reflects a diversity in the objects which constitute the domain. A substantive theory, a theory of the psychological reality which underlies our talk of emotions, must consist of several divergent theoretical strands if it is to be empirically adequate. In addition, of course, it must supply an account which unifies these strands. This can be done either by showing how the various entities which make up the domain are related to one another in a way that justifies classing them together, or by explaining why it would be natural for folk theory to mistakenly classify them together. I shall offer an account which falls somewhere between these two extremes. First, however, I will summarize the shortcomings of the three theories considered in previous chapters.

Shortcomings of Constructionism

In chapter four I gave a detailed account of constructionism, and concluded that this theory cannot give an adequate account of inappropriate emotions, or of the phenomena associated with the psychoevolutionary and affect-program theories. Furthermore, to understand constructed phenomena it is necessary to see that they mimic other, non-constructed phenomena.
Chapter Five: A Multi-Vocal Theory of Emotion

Shortcomings of Cognitivism

In chapter two we listed six prima facie problems for a straightforward cognitivist theory, like that of Solomon, which reduces emotions to sets of evaluative beliefs. More sophisticated cognitive theories were able to meet some of these objections, but three of them appeared to retain their full force against even the most sophisticated. These were the problem of emotions via imagination, the problem of reflex emotions and the problem of physiological response.

Stocker has argued convincingly that we can be pushed into full-blown emotions by imagining suitable objects. Here, the beliefs and desires of the cognitivist analysis are ones we explicitly don't have. The cognitivist can reply that in all such cases we merely imagine we are having the emotion, but when the two opposing interpretations are spelt out for particular cases, Stocker's seems the more plausible. Stocker's own solution, I argued, is irremediably mysterious.

Reflex emotions are emotive responses which occur with extreme rapidity and often in a way which seems to conflict with the beliefs of the subject. Even when such a response is backed up by appropriate beliefs about its object, as when someone is frightened by a loud noise which they later come to believe is an approaching danger, the response can occur so rapidly that it is implausible to regard it as resulting from the usual cognitive processes of belief formation and evaluation. When the response occurs in direct conflict with the subjects beliefs, as occurs when we are frightened by things we know to be harmless, it becomes still clearer that the response is not the result of an evaluative belief in the ordinary sense.

The final problem, that of physiological response, is posed by the same data that was so hostile to the constructivist theory. Certain emotions seem to occur across all cultures. They are characterised by universal patterns of facial expression and autonomic nervous system response, and occur as responses to functionally equivalent stimuli, such as danger, novelty and need satisfaction, in utterly different
cultural contexts. Moreover, the nature of the facial response is to a large extent an arbitrary one. Many facial expressions might be used to express fear, yet the same occurs everywhere. Finally, these responses, and the capacity to recognize them, seem to occur very early in infancy, and in the deaf and blind born, as well as in normal children. It follows fairly clearly that when people ascribe, say, fear, they have a basis for recognizing this state even if the response occurs in defiance of the propositional attitudes of the subject. Conversely, they have a good basis for withholding the ascription of fear, even when such states as the belief in danger and wish to avoid it are present. Thus, even folk practice recognizes that not all emotions are identical with propositional attitude clusters.

Shortcomings of the Psychoevolutionary & Affect-Program Theories

The psychoevolutionary theory of emotion stresses the adaptive nature of emotional responses, and the continuity between our responses and those of other species. Above all, it is intended to be a theory of emotional responses which have a substantial innate component. It is not well suited to highly malleable emotional responses, such as those whose occurrence is limited to a particular culture. Examples include romantic love, melancholy, and the famously "untranslatable" Japanese "amae". Averill (1980b) takes the fairly plausible position that such states involve the more basic biological responses, but as one element among many. This seems quite plausible. The melancholic Elizabethan may have been suffering from a neurohumoural imbalance of some description, but that will hardly explain the extreme cultural specificity of the behaviours which accompany it. This, like the behaviour of the Gururumba when "bitten by ghosts", is something learnt. Melancholia and romantic love are the sort of cases which give constructionism its bite.
The psychoevolutionary theory is usually linked to an affect-program theory of emotional responses. However, the two need not always go together, and the shortcomings of the affect-program theory are not quite the same as the shortcomings of the psychoevolutionary theory. An affect-program is a neural structure that co-ordinates the complex set of events that constitutes an emotional response. It causes facial expression, skeletal/muscular changes, vocal chord changes, autonomic nervous system changes, cognitive changes (such as attention), and perhaps also sensations, to occur in a co-ordinated fashion in response to a stimulus. Affect-programs are linked to a modular system of stimulus evaluation. This appraisal mechanism may be innately sensitive to certain features, but seems to be largely fed by the organism's learning history. I noted in chapter three that the elements of the affect-program response itself could be entirely innate, entirely learnt, or any combination of the two. The innateness of a response offers some support for the thesis that it is mediated by an affect-program mechanism, as the problem of how the response is genetically transmitted can be solved by supposing there to be a dedicated, hard-wired module subserving that behaviour. The problems of supposing that innate behaviours are subserved by a flexible, general purpose cognitive mechanism are discussed later in this chapter, and in appendix II.

Although the innateness of a response provides support for the view that there is an affect-program for that response, the converse is not true. The learnt nature of a response does not rule out the idea that it is subserved by an affect-program. There are, after all, many complex learnt behaviour sequences, such as those involved in driving a car, or throwing a discus, which appear to be highly automated despite being learnt, and even continuing to be malleable if exposed to certain kinds of learning procedures. The learnt nature of a response is evidence against a psychoevolutionary account, but not against an affect-program account.

There are, however, independent reasons for not positing affect-programs for all emotions. First, the behavioural sequences associated with many emotional
states are insufficiently stereotyped. Hope and envy, for example, are not known to be accompanied by any clear pan-cultural facial expressions or patterns of autonomic arousal. Secondly, it may not be necessary to display any overt behaviour in order to (occurrently) have these emotions. Perhaps we feel happier to use the word "emotion" where hope is accompanied by various physiological disturbances (Lyons 1980), but it still seems plausible to assert that someone can "hope" simply by looking forward to a desirable but uncertain event, and this specification mentions only cognitive factors. Likewise, though envy is a complex state or group of states, we are sometimes justified in calling someone envious for no other reason than that they strongly desire that a person more blessed than themselves should lose that blessing, when this is of no material advantage of themselves, and of no moral significance. The folk categories may not admit of a cognitive analysis, but there are instances which are emotions in a clear, pre-theoretical, sense, and which can be analysed purely cognitively. Any behaviour which results from such states is more plausibly explained by the usual processes of action rationalisation by beliefs and desires than by postulating a special mechanism. These emotional states fit the cognitivist theory, better than the affect-program theory.\footnote{These remarks should not be taken to rule out the possibility that the learnt behaviour associated with occurrences of some of these emotions may become automated. All sorts of learnt human behaviour seems to be able to be automated and initiated as a single act with sufficient training. But the differences between such a case and the fully-fledged affect-program responses described in chapter three are more striking than their similarities. The affect program responses are hard-wired, resistant to further conditioning, pan-cultural, and linked to an informationally encapsulated appraisal mechanism. All the two phenomena have in common is their automation. Furthermore, I see little reason to suppose that such responses would be localised in the same, phylogenetically ancient part of the brain, or have a great deal in common at the neural level.}
5.2 The Multi-Vocal Theory

The multi-vocal theory denies that "emotion" denotes a kind of interest to psychology. This claim has been made on previous occasions, by psychologists who have wished to replace the concept of emotion with theoretical variables, such as "arousal". Our thesis, however, is slightly different from their's. A useful analogy would be to compare the folk-concept of emotion to the folk concept of heat. Heat turned out to be a complex phenomena, derived from our sensory experience of such genuine properties as temperature (mean molecular kinetic energy), quantity of heat (total molecular kinetic energy), and conductivity (a matter of the relative ease with which electrons are detached from atoms of different elements). An understanding of the folk concept of heat involves all these factors. What is more, the folk concept is less a useful amalgam of these factors than a misleading conflation of them! Using the folk concept we have no means of explaining how a person can touch an very hot piece of aluminium foil and remain unharmed (insufficient quantity of heat). Similarly, it must have seemed to Wenceslaus's peasant gathering winter fuel that the handle of his axe was warmer than the blade (simply because wood is less conductive than metal). The folk concept is a liability when it comes to understanding the world. It is of (limited) utility only in certain common environments.

Similarly, emotion turns out to be a complex phenomena, involving psychoevolved adaptive responses, cognitive factors and cultural factors. All of these are needed to give an adequate account of what underlies folk discourse about emotion. It is also possible, as we shall see below, to give an account of the origin, and limited utility, of the folk taxonomy, much as we can give an account of the origin, and limited utility, of the folk conception of heat. My response to this is to conclude that emotion and the folk categories of emotion, are not kinds that will be of interest to a developed science. For the purposes of psychology and the philosophy of mind, we should go over to the new taxonomy.
Some later-day Cardinal Bellarmine, however, may be more conservative. Concerned to preserve the traditional folk-categories of emotion, he might try to reinterpret the data we have considered as follows: All we have shown is that instances of the traditional categories are multi-faceted events. We can retain "fear" as a category by considering instances of fear as potentially tripartite events. Those instances of fear where all three elements are manifested are fully fledged instances, while in others only one or two aspects are manifested. By this means we can preserve the traditional folk-taxonomy of emotions. Alternatively, an emotion functionalist might argue that "being afraid" is a highly abstract functional property, which can be realised by any of the sorts of phenomena we have described. The trouble with both suggestions is that they seem quite unconcerned with developing a substantial psychological theory of emotions. Emotions taxonomised in the manner described just don't enter into the generalisations of any useful theory. Hand-waving at our supposed folk psychological competence in understanding emotions with the traditional categories just isn't good enough. For a start, it's not clear that ordinary people do succeed in understanding the emotions with their traditional categories. More importantly, to the extent that these matters are discussed, the ordinary person characteristically alludes to the many different kinds of "love", "fear", etc, for which they have no particular name. The area cries out for a more effective taxonomic treatment. It is as if someone were to respond to the fact that whales aren't fish by saying that all the data shows is that "fish" is a disjunctive kind, or a highly abstract, functional property of many sorts of creatures.

We should not reject out of hand the idea that events consisting of co-occurrences of more than one of the phenomenon described in our three theories may be important event types in themselves. What we must reject is the wholesale preservation of the traditional categories by a taxonomic trick that constructs categories with no justification from theory. Our emotion categories must be framed

---

1 See my discussion of the work of Leventhal (Chapter three).
with the intention of conducing to the formation of lawlike generalisations and broad theories. Theory must drive taxonomy, and not vice-versa. Our three basic categories meet this prescription, as should become evident in the rest of this chapter.

Multi-vocality is important both at the level of emotion tokens, and at the level of emotion types. A token folk emotion ascription may refer to an event of any one of our three types, or to a collection of events of different types. A folk type of emotion, such as fear, may consist of events from any one of our categories, or sets of events from more than one category. Such sets come to be associated in a manner to be described in the next section. This mechanism of association will also shed some light on the co-incidence of different kinds of response in one emotion token.

5.3 The Basis of the Folk-Taxonomy.

The multi-vocal theory should explain the fact that folk-psychology groups such different entities together, both in token ascriptions and in its categories of emotion. This could be done either by providing a real basis for those categorisations, or by showing how folk-psychology would be mislead into postulating them. In fact, I shall do something somewhere between those two extremes. I will show that the categorisations are possible within the general, content based taxonomy of propositional attitude psychology, but I will also show that those categorisations are misleading.

It is possible to taxonomise collections of events of all three types together, either in an emotion token, or in an emotion type, because all three types can be viewed as assessments of the functional significance of the environment to the organism. Thus, all can be associated with states of affairs, and treated as primarily
propositional. Let us start with the central case, that of cognitive emotions. The appeal of a simple cognitive theory is that it identifies say, fear, with the judgement that the environment is dangerous. More sophisticated accounts add other belief clauses, and, importantly, appropriate desire structures, to the conditions for having the emotion. Nevertheless, fear is the cluster of beliefs and desires with which you confront dangerous things, as a result of normal processes of belief formation. So in a quite straightforward sense, fear is the result of an assessment of the environment as dangerous.

Affect-program states can also be related to assessments of the significance of the environment. The affect-program is designed as a response to a functionally defined class of stimuli, say, dangers. The role of the appraisal mechanism is to associate this functional category with an appropriate range of transduceable sensory stimuli, such as loud noises or things with the visual properties of snakes. So there is a sense in which we can say that affect-program fear is the result of an assessment of the environment as dangerous.

There is also a sense in which some constructed emotions correspond to assessments of the significance of the environment. A person has a socially constructed emotion if they behave in the socially prescribed manner when presented with particular situation. Because of the imitative nature of c-emotions, this behaviour will often be a stereotyped version of the behaviour which could be explained by attributing cognitive emotions. Take, for example, the display of fear behaviour in a ritual context where social convention demands fear, but where a little experience reveals that the danger is imaginary. As we discussed in chapter four, people in such situations have sincere c-emotions if they display the appropriate behaviour, but are not self-reflective enough to realise that they are only behaving in this way to live up to a social model, or that they are deliberately putting on an imitation of fear. Where we have such a c-emotion, the person will behave as if they had judged the environment to be dangerous. Constructed
emotions may thus be related to judgements of the significance of the environment, even though the person to whom they are ascribed has never made that judgement.

Some c-emotions cannot correspond to judgements about the environment, but even these may still seem to. We saw in chapter four that there is at least some reason to believe in constructed emotion types. These folk-categories of emotion actually denote nothing but conventional responses. No real mental states type corresponds to this folk category, any more than there is a real kind of ghost possession underlying the behaviour of the Gururumba (see chapter four). Thus, perhaps, "being in love" may be a conventional device to permit certain otherwise unacceptable behaviour, rather than a specific state of mind. In such a case, the emotion term names a fictitious entity, or, on other accounts of reference failure, names the diverse arrays of mental states that might motivate the use of the "in love" device. This would explain the difficulties encountered when trying to describe the contents of such states, or to analytically reduce them to beliefs and desires. But we could still form sentences like "X is loveable", thus creating a fictitious "judgement" that is supposedly made of the object of love by the lover, and, of course, a fictitious property of the object of love. The content of the emotion is that the object of love has this property.

The upshot of all this is that a folk-psychology which types emotions using propositional attitude ascriptions can be constructed. Treating fear as the judgement that something is dangerous will work well for the explanation and prediction of behaviour. But in other cases, although we can treat other emotions as judgements, this will be highly misleading. Tracking down the property that people are judging to be present when they "love," or are "melancholy," will not be possible. Furthermore, where we do have a real property, people may deny that they judge it to be present when they have the emotion. We have gone into these cases at some length in chapters two and three. To take one simple example, it is possible for someone to be frightened of a stimulus which they know to be quite harmless.
Earthworms, for instance. Folk-psychology accommodates this by ad-hoc additions. We incorporate the categories of irrational fear, or its opposite, courage, just as the Ptolemaic astronomers incorporated deviations in the movement of the planets from their perfect paths. Perhaps an exceptionally sensitive person can perform the work of Tycho Brahe and, by attributing enough folk-psychological epicycles, account for all the deviations!

The danger of taxonomising emotions by reference to their supposed cognitive content, is that it leads us to treat all emotions as propositional attitude clusters. But this won't give a good substantial account of what is going on psychologically when we attribute an emotion to someone. The results which show the empirical inadequacy of cognitivism have occupied us in the previous three chapters. What these results show is that the cognitive taxonomy misses the true generalisations covering the various subdomains of emotional phenomena. Even the theory of cognitive emotions will make little progress as long as it is deformed in an attempt to account for affect-program and constructed phenomena. To get a real psychology of emotional phenomena we will have to use taxonomies derived from effective theories, such as the theory of psychoevolution, the theory of modular input systems, and, in the case of c-emotions, some yet to be developed anthropological or sociological science.

I said above that an understanding of the folk/cognitive taxonomy of emotion would shed light on the way in which our three kinds of response could be combined, either in an emotion token, or in an emotion type. In the case of types, the connection is simple. Because folk discourse uses a largely cognitive emotion taxonomy, with ad hoc additions to cover the cases where emotion and cognition come apart, its categories cover many different states which are related to some state of affairs, even if they are related to it in entirely different ways. Thus, "fear" covers a modular response to a stimuli resembling ones that were harmful in the past, a set of propositional attitudes, created by an intelligent assessment of the
threat something poses, and a conventional response to a class of objects which society prescribes as objects of fear.

In the case of a token emotion, a suitable stimulus array may cause more than one kind of response, because it may impinge on us in a number of different ways. Simple stimulus features may trigger an affect-program, while further information leads to the formation of a cognitive emotion, and the recognition of what is expected of one in this sort of situation may cause us to hide, exaggerate or conventionalise the resulting behaviour.

We have been talking as if the folk taxonomy and the cognitive taxonomy are identical. But we saw in chapter two that this is not the case. While the folk taxonomy is content-based, it uses content distinctions selectively. The folk categories of emotion correspond to broad, vague, schemas, which fit many different collections of cognitive states. Fear is the belief that danger is present and the desire to avoid it, but this may be realised either as the belief that a lion will eat you and the desire to run, or the belief that the headmaster will beat you, and the desire to bicycle. What is more, many sets of cognitive states don't make it as emotions, and there appears to be no reason for this in the purely cognitive domain. To really understand the basis of the folk taxonomy we need to recognize the paradigm status of affect-program phenomena in our folk conception of emotion.

5.4. The Paradigm Status of the Affect-Programs

The affect-program phenomena constitute a model of the emotional, or passionate. They are sources of motivation not integrated into the system of rational action on the basis of beliefs and desires. They are involuntary sources of additional motivation. This notion of an isolatable motivational state lies at the heart of the folk-conception of emotion. The characteristic properties of the affect-program
Chapter Five: A Multi-Vocal Theory of Emotion

states, their informational encapsulation, and their involuntariness, necessitate the introduction of a category separate from our categories of belief and desire.

Cognitive emotions come to be regarded as emotions in two ways. First, where an emotion type has some purely cognitive and some affect-program instances, the extension of the emotion term is explained by the fact that in the normal case, central belief formation and affect program triggering run side by side. Since these cognitive contents come, understandably, to be regarded as part of the emotional response, the emotion term is applied to them even when they occur in isolation. If folk psychology works with a largely content based classification of states, as I have suggested, then this extension of the use of the term is almost inevitable. In these cases, the particular types of affect-program response from which the emotion category derives, explains the way in which folk psychology selectively parcels up cognitive states, and labels some of these parcels "emotions".

Emotion types which don't correspond to known types of affect-programs are more puzzling. Jealousy, hope and envy, all of which which may be such types, seem to be based on more or less common arrangements of desires. Working with the idea that an emotion is conceived of by the folk as an isolated source of motivation, we can perhaps say that it is the common occurrence and relative isolation of these sets of desires that leads us to class them with the affect-programs. Furthermore, the occurrence of such strong desires for particular outcomes is relatively involuntary, and disruptive of our pursuit of more general, long term projects, in much the same way as the occurrence of affect-program responses.

It might seem that these states are just what the cognitivist originally promised us. A purely cognitive specification picks out an emotion type. Even in this case, however, the cognitivist is unable to provide us with much of a theory of these states. We want, above all, to know why it is these sets of desires occur so commonly, and not some of the myriad others that the cognitive taxonomy makes
room for. I suggest in the next section that psychoevolutionary factors may be at work in creating common patterns of human motivation such as these. This idea, however, is highly speculative. If there are largely cognitive emotion types that are culturally specific, there will, no doubt, be a very different etiological story to be told about these.

Constructed emotions are modeled upon other emotions, as those emotions are viewed by certain societies. As noted in chapter four, it is the special status which emotions have in virtue of being thought involuntary that c-emotions seek to share. Token c-emotions are modeled upon emotion types whose other instances are either affect-program states or cognitive states of the kinds just discussed. Type c-emotions, if any exist, mimic the general features of affect-program phenomena, and perhaps of those cognitive states that resemble affect-program states in their motivational isolation and involuntariness. Such constructed emotions purport to be urgent, involuntary aberrations from the usual patterns of action rationalised by beliefs and stable desires.

The Future of the Category of Emotion.

So far, I have suggested that the category of emotion will have no role in a future understanding of the mind/brain. I have suggested that it will be replaced by the various alternative categories I have outlined. Another possibility, however, is that the category will be radically redefined. Since the affect-programs are the mental states which correspond most closely to our traditional conception of an emotion, the term may be reserved for these phenomena. The way the term is deployed in some of the works of affect-program theorists such as Ekman suggests that this is what is already happening.

Although such a redefinition would have as its justification the fact that the affect-programs correspond most closely to the folk conception, such a move would exclude the sorts of desire structures that I have hypothesised may underlie jealousy
and envy. These structures are part of the system of rational action guidance by beliefs and desires. If this system is retained in some recognisable way in our science of the mind/brain, these states will be regarded as radically different in kind from the affect-program phenomena.

5.5 Alternative Taxonomies

I have discussed three theories, each of which provides a taxonomy of the states in some subset of the domain of emotional phenomena. I have suggested that these three taxonomies combine to give a close to exhaustive taxonomy of the whole domain. Here I shall discuss the nature of these taxonomies in more detail. Furthermore, emotions enter into a number of other important areas of theory, and this leads us to taxonomise them in a number of alternative ways. I want to examine these alternative taxonomies, not only because of their intrinsic interest, but also to show the connections between our theories of emotion, and theories in neighbouring disciplines.

Taxonomies for Affect-program Phenomena

Affect programs are usually conceived as neural structures, acting to coordinate a complex body of responses. These neural structures have usually been postulated to explain systematic co-occurrence of certain behaviours, rather than discovered through neurological investigation per se. But there has been a certain amount of direct investigation of the neural mechanisms. While detailed accounts of the neural pathways are varied, there seems to be general agreement in locating affect programs in the limbic system, and particularly the hypothalamus (See 3.5). Clearly, affect-programs could be taxonomised at various levels. They could be taxonomised on the basis of their surface features, being treated primarily as patterns of response. They could be taxonomised as particular circuits of neurons,
Chapter Five: A Multi-Vocal Theory of Emotion

or they could be analysed at an intermediate level, abstracting away from any particular neural realisation, but remaining below the behavioural level.

Consider behavioural level taxonomies. The strictest taxonomy of affect-programs at the behavioural level would be species-specific, since affect-programs involve detailed physiological changes, such as patterns of contraction of the facial musculature, and of ANS activity. The species-specific nature of the taxonomy would derive from the species-specific nature of the structures through which the program is expressed. Even if the chimpanzee and man use closely corresponding facial muscle groups to register surprise, there will be considerable physical differences between the surface appearance of the two responses, because of differences in the facial musculature. A looser taxonomy at this level, however, might create categories which run across species. Darwin (1872) uses such a taxonomy when he treats an utilitarian behavioural adaptation in one species as type identical with an expressive adaptation in another. Consider, for example, a particular type of tooth-baring which is primarily a preparation for attack in a species of monkey, but has become almost purely an expressive device in man.

These looser, behavioural level taxonomies involve us in the disputes between the cladists and other taxonomic schools in biology; disputes over the proper relationship between the taxonomy of species and speculations about the evolutionary history of those species. Perhaps the part of these disputes most relevant to the present case is the question of whether our taxonomy should embody a hypothesis about the evolution of the response, or be based purely on the resemblance of the patterns of response. The suggestion that the tooth-bearings of two species are instances of one and the same response might be based on the striking similarity of the two responses, but might also be influenced by hypotheses about the descent of the two species. It may be argued that if we want a taxonomy of affect-programs to provide evidence for evolutionary hypotheses, by displaying appropriate patterns of relationship, it had better not be constructed with those
hypotheses as input. It would not be appropriate to attempt to settle the cladism dispute here, nor is it necessary. It is enough to note that different purposes within evolutionary theory, and perhaps different views about that theory, may require different taxonomies of affect-programs.

Now consider taxonomies at the neural level. Two important factors would determine the appearance of a neural taxonomy of affect-programs. The first is the degree of plasticity in the neural realisation of the programs in different members of the same species. The fact that the structure is innate favours the hypothesis that it is similar in all members of the species. The second factor is the degree of resemblance between the relevant portions of the CNS in different, but allied species. If both human and chimpanzee anger are realised by circuits in the hypothalamus, then, since they lead to interestingly similar outputs, the possibility arises that the circuits may be very similar. Since the responses are unlikely to be strictly identical, even in such close cases, and since not all the differences are merely due to the different shape of the external musculature, an actual identity is not likely. There may, however, be sufficient resemblances for looser taxonomies at the neural level to classify affect-programs in ways that cut across species barriers. This would raise much the same issues as are raised by this possibility at the behavioural level.

Finally, consider taxonomies at intermediate levels. We can describe the programs instantiated by the neural circuits for the affect-programs at various levels of abstraction, and taxonomise the affect-programs at this level. The possibility of cross-species identity exists here, as it does at the neural level. These taxonomies would be of particular interest if the neural realisation of affect-programs were more

---

1 Whether or not a strictly phylogenetic taxonomy of affect-programs is required, such a taxonomy is certainly possible. In such a taxonomy, there may be two fear responses, one common to all members of a genus A, and another to those of genus B, even if fear in species A1 resembles fear in B1 more closely than it resembles fear in any An. In other words, posulated descent would take priority over resemblance. If nothing else, such a taxonomy would be of use in exhibiting a theory of the evolution of particular responses.
plastic than I have suggested. Suppose the same response can be produced as the result of the same genetic factors, but by many different precise arrangements of neurons. In this case the most useful characterisation of the response for the purposes of the neurosciences may be at some intermediate level. This characterisation might, for example, allow us to investigate the role of the genes in the ontogeny of the affect-program system.

In addition to the taxonomies just discussed, affect programs may be typed by their function relative to the broad survival concerns of the organism. Plutchik (1962) places a great deal of weight on this type of taxonomy in his general psychoevolutionary theory. Plutchik relates basic emotions to basic adaptive reactions, such as the need to orient to stimuli, to reject noxious stimuli, to react differentially to overwhelming and confrontable dangers/challenges, and, in higher social animals, the need to integrate and re-integrate into social groupings. He relates such need states to affect programs for surprise, disgust, fear, anger, joy and sadness respectively. These are viewed as evolved adaptive response syndromes. Such broad categories, applying across species, and even across larger groupings, may have a place in a general ecological theory. For example, sadness, in some theories, has a postulated role in re-integration which generalises across all higher social species. Since the functional taxonomy abstracts away from the actual physical nature of the response, the various "levels of analysis" outlined above don't seem to be relevant here.

There are, it seems, three main ways in which affect-program's are to be taxonomised. These are resemblance based taxonomies, phylogenetic taxonomies, and functional taxonomies. The former two can be conducted at various levels of analysis — behavioural, neural and intermediate levels. Taxonomy of all these kinds can be conducted strictly or loosely, to give finer or larger grained pictures of the taxonomic space. This mass of taxonomies may seem confusing, but most of them are used in the scientific literature. We can simplify the picture in at least one
way. It seems to be generally assumed in the literature that taxonomies at all levels from behavioural to neural will be isomorphic to one another. The distinction between the levels will not be critical unless neural equipotentiality problems arise, that is, unless higher level categories are multiply realisable at the neural level. If such problems arise, there will be no taxonomies of affect-programs at the neural level, except taxonomies read off from taxonomies at higher levels whose kind classifications will not be independently motivated at their own level. At present, however, with most experimental work presuming that affect-programs are hardwired at birth, we can ignore the distinction of levels when reading the scientific literature. This leaves us with three taxonomic schemes, resemblance, phylogenetic and functional. A detailed development of the psychoevolutionary theory of emotions, in the style of Darwin’s 1872 book, would probably require the settlement of disputes analogous to those in the realm of species taxonomy, to determine the respective roles of the resemblance and phylogenetic taxonomies. The functional taxonomy finds it’s main use in attempts to develop general theories of the ecological role of emotions. As I argued in section 3.6, the categories of such a theory must be derived from the empirical study of the emotions of actual species if they are to be sound.

In the rest of this section I try to develop a picture of the taxonomies that may be needed in a developed theory of cognitive and constructed emotions. Unfortunately, these remarks will have to be grotesquely speculative. The affect-program theory is the only scientific theory of emotion to be developed in any detail, and furthermore, the sciences to which it is directly related are better developed than those to which the other theories of emotion are related. I will relate cognitive emotions to sociobiology and sociology, and constructed emotions to sociology and anthropology, none of which allow as concrete a discussion as has been possible with the affect-programs because of their relation to evolutionary theory.
Taxonomies for Cognitive Emotions

Cognitive emotions, as conceived in the multi-vocal theory, are collections of propositional attitudes. I discussed in section 1.2 various threats to the enterprise of propositional attitude psychology, and concluded that in the absence of any developed alternative we are forced to use this traditional model to analyse a portion of emotional phenomena unamenable to either of our other theories. In accordance with this decision we must suppose that the basic taxonomy of cognitive emotions is content based. Cognitive emotions will be typed by the contents of the propositional attitudes which make them up, and on this basis, enter into the predictive and explanatory activities of propositional attitude psychology. However, we saw in section 2.5 that the cognitivist is forced to talk in relatively loose terms if he wants to capture something like the traditional idea of an emotion. In giving a cognitive specification of fear, we don't want a taxonomy so fine-grained as to differentiate the state of thinking x is a man eater and wanting to run, from that of thinking x is armed and wanting to cycle away. The cognitive state of fear must be a broad schema, fitted by a range of specific propositional attitudes which, in the appropriate background, count as thinking something dangerous, and wanting to avoid it.

The cognitive taxonomy therefore leaves open an important question. It doesn't explain why certain of these broad schemas are prominent in folk-discourse while others are not. We could type far more clusters than folk-discourse gives us, but only certain clusters seem to be prominent. Perhaps this gap can be closed by pointing to the facts in the face of which emotions are formed. Commonalities between the environments which confront people may explain why these clusters occur in such different social contexts. But it is not obvious that this is so, and in the light of this, we must consider other ways of closing the gap. One obvious possibility is that psychoevolutionary factors may have a role in determining why the same clusters are found in many different circumstances. We have already
Chapter Five: A Multi-Vocal Theory of Emotion

mentioned that where an emotion type has cognitive and affect-program instances, the origins of the category can be explained by the affect-program elements alone. Purely cognitive emotion types, however, proved more interesting. We saw that some of them might be based on common desire structures that intrude disruptively on our pursuit of longer term goals. The occurrence of these desires might have a psychoevolutionary explanation.

It seems not implausible that a propensity to experience such states as envy, gratitude, hope and jealousy may be of adaptive significance for a creature. I have suggested that these states are sets of propositional attitudes that fall under some broad schema. What we are trying to explain evolutionarily is why clusters of this sort should be characteristic of human mental life, rather than many other, equally possible, sorts of clusters. Given the essentially social nature of many of the states just mentioned too, we might expect to find such attempted explanations in sociobiology. Such attempts are, indeed, forthcoming. While not explicitly a contribution to sociobiology, Averill et al's theory of grief (1968) is certainly so in spirit. Weinrich (1980) offers a useful survey of references to emotion in the works of the major sociobiologists. Most of these references are fairly offhand, and occur in connection with classic topics in sociobiology, such as sexual behaviour and altruism. Trivers (1971), for example, is cited as having hypothesised evolutionary bases for moralistic emotions such as gratitude, guilt and trust, as part of the basis of reciprocal altruism. In this context what is supposed to have evolved are cognitive/behavioural dispositions. A propensity to form one sort of cluster of beliefs and desires when treated in a particular way is supposed to be more selectionally advantageous than forming others. Thus, gratitude in response to a favour may be a more advantageous response than indifference, since grateful people are more likely to be the recipients of favours in a system of continuing social interactions. Gratitude in this context must be thought of as a disposition to bestow favours on our benefactor, and to behave in such a way as to assure our
benefactor that this disposition has been formed. In intelligent species, the sort of complex behaviour that is involved in returning favours will be subserved by rational action in accordance with beliefs and desires, and so it is that system that will have to be suitably modified if evolution is to encourage these advantageous dispositions. The sort of story on offer in sociobiology certainly seems to do the explanatory work that it required. It will explain the occurrence of certain patterns of attitude formation, which will in turn explain the cross-cultural prominence of certain broad propositional attitude schemas in folk discourse.

Whether these sociobiological explanations are actually any good is very unclear. Firstly, the whole idea of a sociobiological explanation of any particular detail of a flexible, general-purpose cognitive system is questionable. Kim Sterelny has presented, in an unpublished paper, a general argument against sociobiological explanations of behaviours subserved by non-modular cognitive processes. I take it that this argument would apply equally to the proximal mental results of those processes, such as the formation of an emotion. I am not entirely convinced by this argument, which I discuss in detail in Appendix II, but I have other worries as well. Foremost among these is a worry about whether such sociobiological stories are necessary. We mentioned above that the prominence of certain clusters may be explained by environmental factors. It doesn't seem implausible to suppose that jealousy is commonplace because resource scarcity and sexual competition are commonplace. It may be argued that, given a basis of shared needs/basic desires, most common cognitive emotions are, in the same way, the results of the ordinary processes of belief formation. Finally, there is no very hard evidence on just how pan-cultural such emotions are. If they exhibit cultural variation, then a different sort of explanation may be required, as I'll discuss below.

If, despite all these difficulties, cognitive emotions do come to be seen as evolved structures, then this opens up some interesting, albeit utterly speculative, taxonomic possibilities. For a start, they will be subject to typing both functionally and
phylogenetically, as described for affect-program states, as well as on the basis of resemblance. Furthermore, we can ask whether affect-program states and cognitive states can be type-identical in such taxonomies. The possibility certainly seems open with the functional taxonomy. The same ecological function might be subserved by an affect-program in one species, and a cognitive state in another.

These considerations even raise the possibility of hybrid affect-program/cognitive states. We have already considered and rejected a general treatment of emotions as hybrid states, but in the context of evolutionary theory there may be a use for a typology which treats a hybrid of an affect program and a set of cognitive dispositions as a single adaptive phenomenon. It is not hard to imagine a situation in which any effective response must be both rapid and unthinking, and followed immediately by flexible and intelligent behaviours.

**Taxonomies for constructed states**

I now want to consider the taxonomy of constructed states. Constructed states are intrinsically fairly vague. They arise from a society's sense of what is appropriate to particular situations. When this is stable enough to produce repeatedly recognisable patterns of behaviour in similar situations it produces states which can be type identified. With more diffuse responses it may only be possible to talk about the general structure of belief and expectation that seems to underlie the varied responses to the situation, and try to discern distinctive local elements in this. One way to type such diffuse responses suggests itself immediately. I argued in chapter four that c-emotions are primarily imitations of other emotional states, and that they derive a good deal of their social function from the attitude they command as "passionate", or involuntary. It should, therefore, be possible to type some c-emotions on the basis of which non-constructed state they imitate. This would yield categories like c-rage or c-joy.
Chapter Five: A Multi-Vocal Theory of Emotion

Such a strategy would have no place, however, when dealing constructed emotion types. With such culture specific responses, perhaps including states such as the Japanese "amae", Elizabethan melancholia and certain stereotypes of romantic love, two important taxonomies suggest themselves. These are interestingly parallel to two of the taxonomies described for evolved states, one being primarily etiological (like the phylogenetic taxonomy), the other functional. An etiological typology of c-emotions would construct a history of the elements in the responses, and the social attitudes which make sense of them. Responses and features of responses could then be typed in accordance with their etiology. (The limits of types delineated in this way would necessarily be vague, and identity would probable not be transitive over long stretches of cultural history.) A functional taxonomy of c-emotions would draw attention to the common social pressures met by different responses in different cultures. Many constructionist theorists view the broad function of supposedly "passive" states such as c-emotions as the relief of the individual from incompatible social pressures by the abnegation of agency in his or her decision. I outlined in chapter four the case of ghost-possession among the Gururumba. Averill (1980b) suggests viewing certain instances of anger in the same way. This suggests a basis for typing c-emotions on the basis of their social function. Unfortunately, the day when there is a sociological theory which can yield such a taxonomy will be many years a-coming.

Constructionist theorists such as Averill countenance the possibility of individual c-emotions, which are deviant from social norms and reflect the particular attitudes of the subject. There are parallel taxonomic possibilities in this area to those discussed for the social case, but it is hard to see that they would be of interest to anyone other than a clinician involved in a case study.

I suggested that it might be possible to extend the functional taxonomy of evolved states onto some cognitive states. It may also be possible to extend the functional taxonomy of constructed states onto some cognitive states, We could do this if
some cognitive states could be seen as serving the same social function as some constructed states. A social need might be met, not by prescribing behaviour, but by the existence of beliefs and values such that the cognitive state formed in the subject by the stimulus is such as to produce the appropriate behaviour. The culturally specific emotional responses of one society towards war and death in battle, might be the result of its background ideology interacting with the normal cognitive emotions. Such states could still be cognitively taxonomised, but they might also be grouped with c-emotions for the purposes of sociological theory.

The variety of emotion types and taxonomic schemes discussed in this chapter can be summarised diagrammatically. In the diagram I ignore the differences between the various taxonomies which are applicable to the same range of objects, such as the resemblance, phylogenetic and functional taxonomies of psychoevolved states.
Figure 5.1. Applicability of Alternative Emotion Taxonomies to the Three Basic Emotion Categories.

- Evolutionary Taxonomies
- Sociological Taxonomies
- Folk and Cognitive Taxonomies
- Folk Taxonomy with Fictitious Attitudes
Chapter Six: The Attribution of Emotions

6.1. Historical Introduction

In previous chapters I have referred to the distinction between a substantive theory of emotions and a theory of emotion attribution. So far we have been concerned to develop a substantive account - a theory of the psychological events which underlie attributions of emotion. But the cognitive tradition in emotion theory has been more concerned with providing a theory of emotion attribution - an account of our epistemic access to those psychological states. This interest is a reflection of more general trends in the philosophy of mind. Both behaviourism and Wittgenstein's criterialism directed philosophical attention away from the internal workings of the mind/brain, and concentrated it on folk practice of mental state attribution. It was against this background that cognitivism developed.

In this chapter I shall try to replace the standard cognitivist account of emotion attribution with an account which reflects the multi-vocal view of the substantial nature of emotions. In this opening section I discuss the standard cognitivist view of attribution, and raise certain difficulties for it. In the remaining three sections I discuss the ways in which each of the three main classes of emotional phenomena discussed in previous chapters might be attributed.

Cognitivism and Emotional Sensations

Traditional "feeling theories" of emotion took it for granted that, just as emotions were ontologically individuated by their conscious quality, so they were recognised on the basis of that quality. John recognises his fear, just as he recognises his pain, by the quality of his experience. This account will clearly be anathema to those, like many cognitivists, trained within a Wittgensteinian framework. Consequently, one
of the main concerns of cognitivist attribution theory is to avoid implicating internal qualitative experiences in our recognition of emotions.

Two strategies are particularly apparent in this attack on sensations. The first is to give an account of the third person attribution of emotions, and to apply this in the first person case. The idea is to show that talk of sensations is not necessary to explain how we attribute emotions. Sensations are supposed to be so philosophically problematic that we will not posit them unless we have to. The second strategy is to construe talk of sensations in such a way as to eliminate all reference to subjective qualities of experience, thus rendering them philosophically unproblematic.

Let us consider the first strategy. The cognitive theory of emotions automatically generates a theory of third person emotion attribution. Cognitive theories identify emotions with complexes of propositional attitudes. It follows that our grounds for attributing emotions are simply the grounds for attributing those attitudes. The theory of emotion attribution gets subsumed by the theory of propositional attitude ascription. We ascribe propositional attitudes to others, on the standard view, as a means of explaining behaviour. In line with this approach, classic philosophical cognitivists, such as Kenny, see the self-ascription of emotions as part of the process of making our actions rationally comprehensible to ourselves.

This view of emotion attribution allowed cognitivists to argue that even if sensations do accompany emotions, they are irrelevant to their identification. If emotions are complexes of propositional attitudes which we attribute to ourselves in order to explain some of our behaviours, it is the content of those attitudes that does the explaining, not any quality of experience that accompanies them. Cognitivists were strengthened in this conviction by the results of Schachter and Singer (1962). These results are supposed to show that the same sort of physiological arousal accompanies all emotions, and that the resultant sensations are interpreted by the
subject as a particular emotion only in the light of his/her perceptions and
background beliefs. These results, however, while widely quoted, are widely
scientifically disputed (Plutchik and Ax 1967, Ekman 1983).1

Both the philosophical and psychological versions of the cognitivist view of first-
person emotion ascription run up against some extremely interesting empirical work
by P.D MacLean.2 In Maclean's work on the neurophysiology of emotion there is a
body of evidence for the existence of "disassociated affects". Disassociated affects
are sensations reported by the subject to be recognisably those experienced in
association with a particular emotion, although they occur when there is no suitable
cause or object for that emotion. MacLean (1952) refers to a number of patients
who experienced a sensation, commonly referred to in such cases as an "aura",
immediately prior to epileptic seizures. One patient, for example, described "a
feeling of sadness and wanting to cry". Others felt what they described as
loneliness and fright. MacLean states that:

"The thing I wish to stress in all these examples is the lack of identification
of the feeling with any specific event or person. In short, it is "feeling" out
of context."

The feelings reported are not all free of cognitive connections. Some patients report
irrational and unconnected convictions occurring with their sensations. One patient
felt fear in conjunction with the deep conviction that there was someone behind

---

1 Cognitivist philosophers seem to think that Schacter and Singer provide straightforward support
for the view that sensations have no role in emotion attribution. In fact, there is a considerable
difference between the two positions. Schacter and Singer take it for granted that bodily sensations
are internal states monitored by an interoceptive sense. They think that your sensations tell you
you have some sort of emotion. What they set out to show is that the sensations are the same for
all emotions, and must be cognitively "labeled". Frankly, using their results to support views on
which talk of sensations is to be construed behaviourally, or criterially, is like arguing that you
were never near the scene of the crime and, anyway, you couldn't find the safe.

2 My attention was first drawn to this work by Dr J Gauss, who makes use of it when discussing
cognitivism in his forthcoming book "Value and Justification."
him. But it will not do in such cases simply to attach the feelings to the cognitive elements that accompany them. In the case just mentioned the subject's sensation of fear became intensified if he turned to see who was behind him, even though this action provided him with evidence that his conviction was illusory. MacLean's findings are in accordance with many other studies on epileptic patients. Electroencephalographic studies led MacLean to relate these patients' experiences to discharges in the frontotemporal portion of the limbic system.

MacLean seems to interpret his talk of qualities of sensation in a fairly robust fashion. Like pre-behaviourist, pre-Wittgensteinian philosophers of mind, he takes seriously the idea that the difference between two states of mind may lie in the subjective quality of experience associated with them, and that the subject may know which state he is in on the basis of this qualitative difference.

Such data does not re-establish the feeling theory in one fell swoop. The robust reading is essential if we are to bring it into conflict with the cognitivist picture. But the data does put the onus on the cognitivist to show that it can be interpreted in a way that doesn't conflict with the view that cognitive emotions are always attributed on behavioural/explanatory grounds. Prima facie we do seem to have evidence that subjects can identify a sensation as one associated with a particular emotion, independent of any considerations of context or the subjects' behaviour. The obvious strategy for the cognitivist is to analyse our talk of "sensations" as some sort of ellipsis for propositional attitude ascriptions. This is the second strategy for attacking the feeling theory, which we mentioned above. The latest example of this approach is the work of Claire Armon-Jones (Armon-Jones 1985, 1986a, 1986b) who mounts an attack on the notion of a subjective quality of experience on the basis of a criterialist account of the language of folk-psychology. I offer a theory of sensations that should sidestep all such objections in Appendix III. But even if some such account is accepted, the data will have forced a significant change in the cognitivist thesis. The thesis that cognitive emotions are always attributed on
behavioural/explanatory grounds is very different from the thesis that emotions can be attributed to explain behaviour and also on the basis of feelings, but that feelings turn out to reduce in some way to propositional attitudes.

So far we have been discussing the cognitivist account of emotion attribution as it stood before our rejection of cognitivism as a general theory of emotion. But the cognitivist account of attribution may be able to survive the demise of the substantive cognitivist theory. The retreating cognitivist may claim that, although cognitivism fails to give a substantive account of emotions, it does provide an account of how they are taxonomised by folk-psychology, and a true account of how we attribute emotions to one another. They are taxonomised by reference to their contents, or supposed contents. Therefore, the cognitivist may argue, we must attribute them on behavioural/explanatory grounds, in much the same way that we are generally supposed to attribute propositional attitudes.

We have already subscribed to part of this view in earlier chapters. Folk discourse does seem to use a cognitive taxonomy. It seems to classify emotions by reference to the states of affairs which elicit them when the system functions normally. The folk think of fear as the response to danger and fearing X much the same thing as believing X to be dangerous. It is precisely because folk discourse uses this cognitive taxonomy that emotions which conflict with the subjects beliefs seem peculiar. But this does not mean that the folk depend upon the cognitivist method of emotion attribution. They frequently encounter cases where the supposed content of the emotions they ascribe contradicts the beliefs they ascribe. These cases arise because cognitivism is not, in fact, true. The folk are able to recognise these cases for what they are because, as the rest of this chapter will show, the diversity of states that underlie attributions of emotion leads to a corresponding diversity of ways in which such attributions can be made. The cognitivist theory of attribution, at least as a general account, will not survive the fall of the substantive cognitivist theory.
6.2 Attributing Cognitive States

In chapter five I outlined the substantive account of emotions which I have named the "multi-vocal" theory of emotion. In this account we do find certain "cognitive emotions", but they are not those of the original cognitive theory of emotions. The new theory does not envisage that folk-categories will be type identical with certain complexes of propositional attitudes. Take a folk category like "fear". Some tokens of fear will, on the new theory, be propositional attitude complexes. Others, however, will be affect-program responses. Others will be constructed phenomena. On yet other occasions, the token state will be a complex of two or three of these types of phenomena. A single component may not be anything that the folk would call an emotion if it were to occur in isolation.

In this section I want to discuss our grounds for emotion attribution in cases where the emotion is a propositional attitude complex, or when recognising the cognitive component in a situation is necessary for the folk-attribution of an emotion. These are our "cognitive emotions". To a large extent, we have backed ourselves into a corner by specifying this class of states in the way that we have. If cognitive emotions are collections of propositional attitudes, then however propositional attitudes are attributed, so must cognitive emotions be. To say this is not so much to answer the question of how cognitive emotions are attributed, but to refer the questioner to an ongoing philosophical debate. We noted in chapter one that the ontic status of propositional attitudes is problematic. There are connected problems about the nature of the explanatory relation between propositional attitudes and behaviour. It may be that propositional attitudes only explain behaviour in some special sense; by "normalising" it within some non-empirically derived system of rationality, for example. But if this is the case, propositional attitudes are unlikely to be interpreted as parts of the mechanism of the mind-brain, and our substantive account of the cognitive emotions will be in trouble too. That area of the domain of emotional phenomena will have to be examined in the light of the new account of
the mind/brain phenomena that underlie them. Our account of the area, and the corresponding account of emotion attribution, will be part of the folk psychology which has been found wanting for the purposes of a scientific psychology.

6.3. Attributing Affect-Program States

In chapter three we discussed the psychoevolutionary theory of emotion, and specifically the modern affect-program theory. We concluded that there is compelling evidence that there are a number of emotional states which occur in all human cultures, and are adaptive responses which we have inherited from our evolutionary forebears. In this section I discuss the attribution of these, "affect-program" states.

Attribution on the Basis of Facial Display

Following Darwin, modern researchers have attempted to show that there are a number of distinct facial expressions which occur in all cultures, even those which have been securely visually isolated until the present day. These expressions occur as responses to functionally equivalent stimuli in these various cultures. They are consistently recognised as responses to those stimuli by members of the various societies. The fear expression, for example, occurs in each society, and is recognised as the response which people have to things which are, in that context, sources of danger. However, even if the propensity to produce that expression is inate, it is a further question whether the ability to recognise the significance of that expression in others is inate. This is an issue of some importance. If the significance of the expressions must be learnt, it seems probable it must be learnt by observing the co-incidence of that expression and situations which can be seen to have a particular significance for the organism, or easily interpretable behaviours,
Chapter Six: The Attribution of Emotions

such as fighting or fleeing. This learning process might provide a basis for claims about the "priority" of behavioural/explanatory modes of emotion attribution.

General evolutionary considerations might lead us to expect that recognition would be innate. It would, after all, be an advantage for an infant to recognise such things as aggression in its conspecifics. Even earlier in life, the affective relationship between the mother and newborn child is of the first importance for the child's survival. On the other hand, given the enormous potential of infants for acquiring such information, and the possibility of some innate predisposition to acquire this information in particular, perhaps infants could acquire the information quickly enough to eliminate any significant advantage. At this point we need to open the possibility of recognition being more or less innate. We should recall the discussion of innatness in chapter three. We saw there that the notion of a trait being innate is best replaced by the notion of a trait's emergence being more or less invariant across the range of possible environments. We referred to this as traits degree of "buffering" against the environment.

Trevarthen (1984) summarises an extensive series of experiments connected with this question. He carried out elaborate observations of mother/child interactions, paying particular attention to the reciprocal influence of the facial expressions of the two parties. By interfering with feedback from mother to child, by interposing one-way mirrors, and playing film of the mother to the child, Trevarthen was able to show even very young infants are acutely sensitive to the appropriateness of the feedback which they receive to their output of expressions. In many cases, sequences of maternal expressions which were entirely satisfactory in their original role as part of a sustained "conversation" elicit distress when they are presented as a fixed sequence. A request to the mother to adopt an unchanging expression also rapidly elicited distress. The evidence suggests that some basic perceptual ability to distinguish emotional expressions, and to treat them as significant features of the environment emerges extraordinarily early. Field et al (1982) report that a sample
Chapter Six: The Attribution of Emotions

of 76 infants with an age around 36 hours appeared capable of imitative responses to emotional facial expressions.

This sensitivity to facial expression contrasts with the general insensitivity of infants to stimuli which would be emotionally charged for adults. The early and uniform emergence of this ability suggests that the ability to recognise facial expressions may have a fairly high degree of innateness.

Attribution on the Basis of Context

There are a number of other ways in which attributions of affect-program states may be made. First, we can attribute affect-program states simply because appropriate stimuli are present in the context of attribution, and so an appropriately linked affect program ought to be functioning. This makes it possible for someone to infer that they are afraid, because they know themselves to be in danger. They can also infer that someone else is afraid from the fact that they know them to be in danger. Given a normal learning history, and normal environment this procedure is perfectly reliable. Since folk-psychology seems to use a cognitive, content-based, taxonomy it seems likely that folk-practice uses such a method, and does not distinguish the affect-program state from cases where fear consists merely in the belief that we are in danger and the wish to avoid the danger. But folk practice recognises the paradoxes which this procedure creates in abnormal cases. In these cases belief formation and the functioning of the affect-program cease to run in tandem. Folk-psychology admits that people can be unafraid when they are trying to avoid a danger, and can be terrified of things they know to be harmless, such as open spaces, or small mammals.

Direct Flow of Information from Affect Program.

Secondly, we may be able to attribute emotions to ourselves simply because there is an information flow from the affect-program system to our central cognitive
processes. The evidence from Ekman et al.'s experiment on Japanese students suggested that there are systems capable of interrupting some aspects of the response almost as soon as they are triggered. Such an event would have to be preceded by the passing of the information that the response is being triggered. If this intervention is mediated by higher cognitive processes, it is possible that we are made conscious of our emotional state automatically, as part of the output of the affect program. However, given the automation of the Japanese response, it is equally possible that this information flow is not open to conscious inspection.

If there is such an input to our consciousness from certain affect programs, this need not imply that we are always know whether we are experiencing the affect-program state. After all, there are very few informational inputs to our higher conscious processes that are not affected by attentional factors. Even pain perception, an unpleasantly automatic process, can be blocked by other urgent cognitive activity.

**Attribution on the Basis of Sensation**

Finally, sensations may have a role in the self- attribution of affect-program states. Perhaps this is just one way in which the possible information flow just discussed may occur. Information about the operation of affect- programs may reach us in the form of sensations. This possibility has been suggested by P.D MacLean, in connection with the results of his referred to in section 6.1.

MacLean (1952) reports that discharges in the limbic system of epileptics lead to sensations which the subjects claim are the sensations they experience in various emotional states. In this and later papers (MacLean 1960, 1969, 1970, 1980) MacLean makes some interesting, if speculative, connections between this data and the localisation of emotions in a phylogenetically ancient portion of the brain. We have already mentioned the connection between these ideas and Fodor's evolutionary account of the origins of modules. MacLean draws attention to the
Chapter Six: The Attribution of Emotions

preservation within the human brain of structures which represent the state of the
brain at earlier stages in its phylogeny. He isolates what he refers to as the reptilian,
paleomammalian and neomammalian brains. In the context of his emotion theory he
argues that the paleomammalian segment, which is largely identical with the limbic
system, supports a level of brain functioning in which sensation is central. There is,
of course, a tradition, going back at least as far as Walter Cannon's work in the
inter-war years of regarding the limbic system, and particularly the hypothalamus,
as the main locus of emotional functioning.

MacLean's ideas provide one picture of how sensations could be important in the
attribution of emotion. They could be the way in which the operation of the affect
programs is made manifest to the conscious mind. But there is a far older tradition
which suggests a different way in which sensations may play a role. The well
known James/Lange theory of the 1890's suggested that our sensational access to
our emotional states was a consequence of two other facts. First, they claimed, the
emotions have characteristic effects on autonomic nervous system activity.
Secondly, feedback from the ANS creates interoceptive sensations, which we
usually associate with the functioning of our internal organs. The ANS feedback
from the characteristic visceral activity associated with emotion is the emotion
sensation. This theory was once considered to have been refuted by W.D Cannon's
work in the inter-war years, which drew attention to the long time delays in visceral
feedback, and to its relatively weak and undifferentiated nature. The Schachter and
Singer experiments, mentioned above, were also thought to tell against the
James/Lange theory. But this dispute is by no means settled. Ekman (1983) for
example, published results indicating differentiation in patterns of response from
various measures of visceral activity, including alterations of heart rate and the
temperature and electrical resistance of the skin.

New theories about the role of feedback have also been developed, which are less
vulnerable to Cannon's data. Cannon's argument that feedback from the viscera is
too slow to account for the immediacy of emotional response is still a good objection to the James/Lange theory, but is not an objection to more recent theories which use visceral feedback to sustain and intensify emotional responses. Prominent amongst these are the work of Tomkins (1962, 1963, 1979, 1980) and Izard (1971, 1977). Both these theorists take the role of feedback from the viscera to be that of sustaining the emotional response via a positive feedback loop. They also give considerable attention to feedback from the faster acting somatic ("voluntary") nervous system, and especially the facial musculature. This is reminiscent of Darwin, who suspected that the internal feeling component of an emotion could be generated by placing the facial muscles in states they would usually occupy in that emotion.

Like Maclean, Tomkins and Izard take the sensational aspect of their theories very seriously. They regard the quality of the sensation as playing an important causal role. Certain qualities of sensation are acceptable or unacceptable to the organism in such a way as to reinforce behaviours which promote acceptable sensations, and prevent unacceptable behaviours. This idea, of course, can be extended to other sensations, such as pain and sexual feelings. Tomkins himself draws the analogy when he tries to explain what he means by saying that certain feedback is inherently "unacceptable":

"If, instead of pain, we always had an orgasm to injury we would be biologically destined to bleed to death." (1980. p147)

Like MacLean, Izard and Tomkins seem to believe that two states may be different because they differ in subjective quality, and that this difference may affect their causal properties. If this is so, and if feedback from the facial musculature is an important part of the experience of emotion, this would allow the subject to judge his emotional state from its subjective quality. If visceral feedback also turns out to
be differentiated amongst the emotions then this may contribute to the same process.

Philosophically, giving causal significance to qualities of sensation is highly controversial. I address this controversy in Appendix III. I believe that we should take the scientists seriously, instead of dismissing their work because of some philosophical dogmas acquired during the reign of behaviourist and criterialist theories of mind. There may be philosophical debate over the interpretation of this talk of "sensations" and "qualities of sensation", but this should not lead us to dismiss theories which use these ideas. Instead, we should make a serious attempt to find a coherent interpretation which can play the appropriate role in those theories.

6.4 Attributing Constructed States

The final class of states referred to by the multi-vocal theory are the constructed emotions, or, as I have elsewhere labeled them for brevity's sake, the c-emotions. C-emotions are learnt patterns of behaviour which reflect and reinforce the values and structures of the societies in which they occur. They are distinguished from mere pretend emotions by their subject not being fully conscious of their constructed nature. We use the idea of a c-emotion to explain human behaviour of an institutionalised or ritualistic nature, which is interpreted by the subject, and/or the society around them, as the occurrence of a natural and involuntary response. We justify calling them emotions by the fact that folk-psychology does not distinguish them from the other emotions, and because many instances of emotional behaviour involve constructed elements intimately entwined with psychoevolutionary and cognitive ones. Most c-emotions seem to be attempts to mimic emotions of other types, but we speculated in chapter four that there may be some which represent independent emotion types, and which mimic the general properties of emotional states as they are conceived in that society.
Knowing how to attribute a c-emotion means knowing the conventions of a particular society. It means being able to recognise a pattern of responses to some stimuli as the pattern prescribed by society, and given a particular emotion label. Only someone who understands Elizabethan convention will recognise the symptoms of Hamlet's melancholy for what they are. The attribution of c-emotions in the third person thus depends primarily upon associating the term for the emotion with a collection of behaviours.

**Self-Attribution of C-Emotions**

The self-attribution of c-emotions may have a different basis. We discussed at some length in chapter three the various ways in which knowledge of the constructed nature of a set of behaviours may be kept from the consciousness of the subject. Self-attribution could, when the level is deep enough, depend upon the recognition of one's own behaviour, as in the third person case. In the case where the c-emotion rests upon a relatively shallow form of self-deceit, however, having the emotion may be something you decide upon, rather than find out about. You may know that your emotion is anger because that's what it's meant to be!

Despite the conventional basis of c-emotions, it is not impossible that they may sometimes be attributed on the basis of characteristic facial expressions or self-attributed on the basis of sensations. Constructionist theorists often refer to the "recruitment" of more biologically basic states as elements in constructed ones. A child who, for example, learns to give a socially prescribed display of ritual terror in some religious context may be able to trigger the affect-program for fear as an element in the display. Given the importance of feedback from emotional expression and behaviour in theories of emotional sensations, this would fit in quite easily.
Mistaking C-States for Cognitive States

Because c-emotions are intended to mimic natural emotions they tend to occur in contexts where they can be fitted into a behavioural/explanatory context and attributed like cognitive emotions. Someone from another culture might interpret the behaviour of the Governor General on a royal visit as displaying the presence of beliefs and desires amounting to the state of being in awe of the queen. This is what allows folk discourse to apply its cognitive taxonomy to constructed states. But the person who attributed this cognitive state to the Governor General would be fundamentally mistaken, since his behaviour is not caused by that set of beliefs and desires, but by social conventions which mimic its results.

Although such a person would be mistaken, at least they are attributing a set of beliefs and desires, which might, in some other case, have been present. It is possible for attributions of constructed emotions to be mistaken in a much more outlandish way. In chapters four and five we mentioned the possibility that there may be constructed emotion types in folk discourse. Using the cognitive taxonomy, the folk would attribute these by attributing a fictitious mental state. Some authors have suggested that "...loves..." is such a fiction. The term either fails to refer, or, on another account of reference, refers to the whole range of ways in which a person might be motivated to display this conventional behaviour. Love, on this account, is a fictional condition, whose attribution and self-attribution serves to excuse certain otherwise deviant patterns of behaviour in a particular culture.

6.5 Summary

In this chapter we have seen a diversity of means of attributing emotions. This diversity reflects the diversity of psychological events which underlie emotion attributions. We have seen that cognitive emotions are attributed in the same way as propositional attitudes, as part of the activity of rational action explanation. The traditional cognitivist view is sometimes the correct one.
The affect-program states, on the other hand, are frequently attributed on the basis of facial expression. We have seen that there is reason to suppose that this reflects an innate ability on our part. We also noted that affect-program states can reliably be attributed to properly functioning individuals in their normal environment on the basis of their having a suitable stimulus presented to them.

We considered various ways in which affect-program states could be self-attributed. It may be that there is a direct flow of information from the affect-program system to our higher cognitive processes, so that we sometimes just come to believe that we are afraid, angry or whatever. We considered the possibility that this flow of information might take the form of an interoceptive sensation. Finally, we considered a range of theories which suggest that our interoceptive awareness of the visceral or muscular changes caused by the affect programme may tell us which emotion we are experiencing. Issues directly concerning the nature of sensation were deferred until Appendix III.

In section 6.4 we considered the attribution of constructed states. We saw that they are attributed to others on the basis of behaviour patterns which have been brought together and given a label by the appropriate culture. We also considered some ways in which they could be self attributed, other than by a reflexive application of this first process. First, it may be the case that we are partially aware of what we are doing when we display a constructed state. In this case, I know what I intend my emotion to be. Secondly, a constructed response may "recruit" as components affect-program responses, and our overall condition might be recognised on the basis of the resultant sensations.

Finally, we noted that in folk-contexts constructed states can be mistaken for cognitive states, and that attributions made on this basis can be quite reliable. Eliciting this confusion is, after all, part of the function of constructed responses.
For expository reasons I have discussed cases in which what underlies the folk attribution is predominantly one or other of our three phenomena. However, the same epistemic routes exist in cases where a set of these phenomena co-occur. The same stimulus array may trigger an affect-program, lead to appropriate beliefs and desires, and to the display of culturally appropriate, learnt behaviour patterns. To the folk, this array of consequences may be one instance of fear, disgust, or whatever. They may use any one, or any set, of the epistemic routes to make their attribution. I do not see that any difficulties are caused by these cases, or that any useful purpose would be served by giving detailed examples.
Chapter Seven: Moods and Character Traits

7.1 Introduction

My strategy in this chapter will be to construct complimentary models of moods at the functional level and at the level of physical realisation. I will show that moods, like emotions, can be thoroughly integrated into a scientifically respectable philosophy of mind. In the process I hope to throw some light upon the nature of traits of character. The account will show that emotions, moods and traits are three radically different types of phenomena.

I do not propose to offer a definition of "mood" prior to stating my theory of what moods are. Instead, here are some prototypes. Common moods include depression, elation, anxiety, irascibility. It is states of this general type with which I shall be concerned here. I do not believe that our pre-theoretical conception of moods provides us with any precise way of delimiting the class of moods. Accordingly I shall be more concerned to provide an account of certain prototypical mood phenomena that hangs together with the rest of our science than to provide an account of all and only what English speakers call moods.

In the next section, I criticise another published account of moods which attempts to work within the structures of a scientifically respectable philosophy of psychology. In section three I describe some historical precursors of my own account. Section four describes my development of these suggestions. In section five I give an account of the use of mood terms in folk discourse which suggests that my theory captures important aspects of the folk conception of moods. This justifies calling it

---

1 As I spell out in Appendix III, I do not accept the functional/physical distinction in its traditional form. What I do in this chapter would be more properly described as giving a model at a fairly high level of description, where I am concerned only with the behaviour of the system, and a model at a lower level of description, where I am concerned with the precise neural mechanisms which produce that behaviour.
a theory of moods. In section six I show how the theory is consistent with, and even suggestive of, the current neuroscience of mood. This allows me to argue for the superiority of our account to other accounts at the functional level, on the grounds that it is more likely to be psychologically realistic.

7.2. Lormand's Theory

I know of only one other attempt to give a model of mood phenomena which can be integrated into a scientifically respectable philosophy of psychology. This is Eric Lormand's theory (Lormand 1985). Lormand is a cognitivist about the emotions, but is struck by three contrasts between emotions, as he pictures them, and moods. He doesn't think moods have intentional objects, he thinks they influence all aspects of your mental life, without respecting content based boundaries, and he thinks explanations of, and involving, moods are usually brute causal, and not intentional. Generally, then, he doesn't want to extend the cognitivist account to moods. He wants to think of them as contentless states.

Lormand builds his theory using a distinction between "latent" and "active" intentional states. Active intentional states are those presently capable of guiding action. Latent states are those which are not presently capable of guiding action, but which are, in some sense, present. Lormand doesn't want to discuss the sense in which they are "present", and seems to assume that there are current accounts of mental representation which will give him what he wants. This suggests that he thinks his distinction will map onto some existing one, such as the distinction between implicitly and explicitly represented information.

Lormand's theory is, simply, that mood variables determine the class of intentional states that are active. The "mood box" is a filter determining what part of our store of intentional states is available to guide our actions. Anecdotally, the idea seems to be that you only remember that your dog is irritating when you're in a bad mood,
and how much you value your relationship with your spouse when your happy or anxious, rather than depressed or irascible.

There are a number of reasons why this proposal doesn't strike me as a good one. First, moods don't only affect intentional action. They have specific physiological effects. They also have direct causal effects on behaviour. In many primates depression is associated with characteristic patterns of huddling, rocking and self-gripping. It is implausible to suppose that this behaviour is rationalised by some previously held beliefs and desires which are activated by the depression. Finally, moods seem to affect behaviours which are plausibly under the control of modular, rather than central systems. Depression can affect the performance of all sorts of mental processes, right down to the reflexes. It's hard to see how any of these phenomena get dealt with by a theory on which moods are a filter on states responsible for intentional action.

Furthermore, why think that the states allegedly caused by moods are ones that existed before the onset of the mood? There's just no reason to believe that someone who is irritated by their dog when in a bad mood must have had a prior implicit belief that the dog is irritating! In fact, it may be characteristic of the states we ascribe to the effects of mood to be out of step with the normal person's way of thinking. A sweet, loveable spouse may become spiteful and unreasonable when acutely depressed, without having implicitly believed their various resentments and accusations beforehand. In general, the effects of moods seem as likely to be the formation of new states as the activation of existing ones.

Our own theory will avoid both these sets of problems, and will have one further, significant advantage over Lormand's. His theory, while a laudable attempt to work within the structures of recent philosophical psychology, pays scant attention to the neuroscientific data. Our own model will be consistent with many of the
known facts about the neural basis of mood phenomena. It will provide a new way of relating many of these results, and suggest forms of future research.

7.3. Moods as Higher-order Dispositions

The notion of an higher order disposition is introduced by C.D Broad, in his "Examination of McTaggart's Philosophy" (Broad 1933). At the first level we have ordinary dispositions, such as a magnet's disposition to attract iron. At the second level we have properties like magnetizability. This is the disposition to acquire the first order disposition of being magnetic. There is no reason to stop at two levels, although that's all we need to deal with the phenomena of magnetism. Broad proposes that the real upper limit on hierarchies of dispositions comes when we reach a supreme disposition. This is a property of a substance which it could not lose without ceasing to be that substance. There are a number of problems associated with this notion, but we will not need to consider them here.

Broad's idea was noticed by the psychologist Vincent Nowlis (Nowlis 1963), who suggested that moods might be distinguished from emotions as higher order dispositions to lower order ones. Moods, he proposed, are dispositions to have emotions. He also suggested that a third order of dispositions in the same hierarchy may be called states of temperament. The picture that emerges from these suggestions is, in its simplest form, that a long term trait like irascibility is the state of being markedly disposed to be in an angry mood. An angry mood is a disposition to get angry easily, and anger itself is a disposition to all sorts of behaviours and mental state changes. To take another example, being a depressive would be possessing a marked disposition to become depressed. Becoming depressed would be an alteration in dispositions to have such emotions as joy and sadness. This proposal has many serious flaws, as we shall see, but it contains the germ of a very illuminating approach.
The simple idea that emotions, moods and traits form a hierarchy of dispositions is inadequate in four main ways:

i) The simple model fails to allow for the direct influence of moods on behaviour. There seems no prima facie reason to deny that someone's bad mood can affect, say, the aggressiveness with which they perform mundane tasks, or their facial expression, without the interposition of some occurrent emotional state such as anger. In fact, there is every reason to assert the contrary. To accommodate this, we must allow direct connections from moods to behaviour.

ii) Moods don't just dispose to emotions which are somehow conceptually "appropriate", as when depression disposes to sadness. The known consequences of depression are wide ranging, and include states such as anger and abnormal sexual excitability. We need a model which allows for these connections.

iii) There is no reason to suppose that all emotions, moods and traits come neatly packaged in sets of three, one of each kind. We should avoid this assumption.

iv) It is at least plausible that emotions may be the causes of moods, or traits. A person may be put into a bad mood by having been angry, or into a good mood by a pleasant surprise. Likewise, frequent exposure to frightening or challenging stimuli, especially during a formative stage of development, may create a trait of timidity or aggressiveness. This is not possible on the traditional model of higher order dispositions, since on that model it is a necessary condition of possessing any contingent, n-order disposition $D$ that the bearer already possess $D^*$, where $D^*$ is the $n+1$ order disposition to have $D$. Lower order dispositions would seem unable to cause higher order ones, since those higher order ones must already be in existence.
7.4. Moods as Higher-Order Functional States

As a first step we need to recognise that moods and traits are complex functional states, not philosophers' dispositions like fragility. Let us therefore introduce the idea of an higher order functional state\(^1\). Suppose a single physical system is capable of realising, at different times, a number of functional descriptions \(F_1, ..., F_n\). Suppose further that the same (physical) system realises another functional description which gives as its outputs the system realising one or other of \(F_1, ..., F_n\). Then we have the basis for a hierarchy of functional states (although, as we shall see below, more is required to justify using the notion of a hierarchy.) Consider by way of example, the computer upon which I am writing this. When I set up for the day the computer realises a functional description. When I provide a certain input, the output is the computer realising a functional description which may, for example, be a word-processing program. In other words, one of the most convenient ways to describe the system, at least at this superficial level, is as a number of systems, nested within one another. Similarly, a person is a system capable of realising a number of functional descriptions. One such description, or more likely a set of them, is our normal functional model of persons. The normal character of a person we are acquainted with is probably an approximation to one of these. The alternative model we have of a person when we know them to be depressed will be a significant deviation from the normal model. We can see our overall understanding of persons as involving a number of functional models nested within a higher level model. That higher level model is our folk-theory of moods and traits of character. The model relates various kinds of inputs to outputs consisting of the person fulfilling one of the various first level functional descriptions within their overall capacity.

\(^1\) Not to be confused with a higher level functional state, a quite different, and better known notion, which I use in section 7.6 and appendix three
Chapter Seven: Moods and Character Traits

We can see higher order philosophers dispositions as trivial cases of higher order functional states. Take the standard example of magnetism. A piece of iron can realise two simple functional descriptions. The first specifies that if certain substances are presented a certain attraction results. The second specifies that no attraction other than normal gravity results. There is a second order description of the piece of iron which states that the input "stroke with a magnet" takes the iron from realising the second description to realising the first, while the input "strike sharply" takes it from the first to the second.

The new model avoids some of the problems enumerated at the end of the last section:

i) Moods can now cause new patterns of behaviour immediately, as well as via the mediation of emotions. A change of mood alters the overall functional description of a person, and as part of this, it may directly bring it about that, under certain circumstances, patterns of behaviour characteristic of that mood under those circumstances will arise.

ii) Similarly, higher order states don't cause a single "conceptually appropriate" lower order state, as when depression causes sadness. The model will allow a diverse range of effects to follow from someone becoming depressed — anger, or sexual voracity, as well as sadness.

iii) For the same reason we do not need to conceive of the emotion/mood/trait system as consisting of discrete hierarchies with an appropriate state at each level. The notion of a discrete hierarchy of states has no place in our model, which consists of a single hierarchy of overall functional descriptions.\footnote{The model I shall discuss here uses overall functional descriptions of the person for simplicity's sake. If our theory of the mind is modular in character, a more correct picture may confuse the hierarchical structure to our model of certain sub-systems. The discussion of moods at the neural level in section six of this chapter, however, would suggest that moods are, if not global phenomena, at least very diffuse, affecting several different areas of the brain simultaneously. This may suggest that their effects are not confined to one module.}
iv) The fourth objection, however, still seems pressing, although in a slightly altered way. The new model does not exclude the possibility that lower-order states could cause higher-order ones. Take the simple computer analogy. There is an input to the word-processing programme that takes the machine back to its original, monitor, state. But if this is the case, what are the criteria for higher-orderedness? Why aren't the mood/trait states just states of the overall functional description of the system, at the same level as all the others?

To get round this worry we need to go back to the idea of an higher-order disposition. We need to appreciate a basic problem with this notion as traditionally applied in philosophy. The sorts of states Broad calls higher-order dispositions aren't distinguished from mere abilities. An ability is a vague dispositional state, one whose eliciting conditions are left in obscurity. Broad is able to infer that a bearer has Dn+1 from the fact that it has Dn, because Dn+1 is just the property of being able to bear Dn! The real higher-order dispositions to Dn, on the other hand, are properties of bearing Dn under eliciting conditions C. There may be many of these corresponding to a given Dn, each relating Dn to some conditions C.

It follows from this that a lower-order state can cause a corresponding higher-order state, even when we're dealing with simple dispositions. It just can't cause the one whose existence it depended on. Once we get away from the pernicious locution of abilities, we can confront the real question of what makes a dispositional, or functional property higher ordered.

A dispositional or functional property D* is one order higher than another, D, if two conditions hold. First, D* must be a disposition to have D under some eliciting conditions C. There is nothing to prevent this situation obtaining symmetrically
between D and D*. Secondly, in the context of some theory, it serves the interests of theory construction to introduce an hierarchical structure amongst certain properties, in virtue the high salience within the theory of certain dispositional relations. It is this second element that gives the direction of the hierarchy. The reason for this salience may be purely pragmatic. Thinking again of our computer analogy, we could, if we wished, represent all the machine's capacities with a single level network of functional states. However, the fact that there is a state $S$ from which we can readily put ourselves into a number of states suited to the performance of particular tasks makes it convenient to use a model in which $S$ is at a level distinct from that of the task-specific states.

I hope to show that there is a more substantial reason for the salience of the dispositional relations on which I have based my theory of moods and traits. The hierarchical structure at the functional level reflects a real feature of the mood/trait system at the neural level.

We now have our theory of moods in its final form. Moods and traits are states with certain highly salient properties for the purposes of our psychological theories. These properties are that they cause global changes in our propensities to occupy certain other states, and to respond in certain ways to certain stimuli. The salience of these effects gives utility to a model in which moods are higher order states which determine which of a range of lower order functional descriptions the person occupies. In the next two sections I shall offer two kinds of evidence for this theory. In section four I shall argue that the theory captures an important aspect of our folk conception of moods and traits. In section five I shall argue that the functional model embodied in our theory should be preferred to a single level functional model, because the hierarchical structure directly reflects the neural realisation of moods and traits.
7.5. Support from the Folk Conception.

The use of moods and traits in folk psychology is something like this. Certain folk generalisations governing states like emotions and desires don't always apply straightforwardly. We have a general conception of the behaviour that would be likely to result when someone wins a national rowing competition, provided that they are an ordinary competitor, and that they win in the ordinary way. But we recognise that our expectations depend on mood and character. Certain moods states, such as depression, may greatly increase the degree of success needed to initiate the normal celebratory response. Such variations in response between persons, and between the same person at different times could, of course, be accounted for by a complication of the picture at the level of emotions, beliefs and desires. But this is often very difficult. Julie's failure to enjoy winning intervarsity sculls could be explained by ceasing to think that she desired to win, but this would force a fundamental reappraisal of her character. We get a much simpler explanation if we suppose that she is depressed. Rather than altering otherwise well grounded belief and desire attributions, we can posit this higher order state which alters the transition probabilities between those states, and the input/output relations we need to account for. Character traits have a similar function, but are particularly employed to account for the variation between individuals. When Julie and her pairing partner learn that they've been disqualified, one gets depressed, the other doesn't. Now, supposing that we have reason to believe that the significance of this event is the same for both of them, we can save ourselves falsifying otherwise well grounded beliefs by the hypothesis that one of them is a depressive, or that one of them is of a buoyant disposition.

What I am proposing is that ascribing moods and character traits serves to account for deviations from a central, "normal" functional model of persons. This suggestion is not new in itself, something like it has been propounded by Adam Morton (1980), and doubtless elsewhere. But if this account of the role of mood
and trait terms in folk discourse is correct, then our hierarchical theory captures an important aspect of that discourse. On our account, folk moods and traits are variables determining which of a range of possible functional descriptions is true of a given person at a given time. In other words, their most salient role in the (folk) theory is as determinants of lower order functional state. If, as we are about to argue, our theory is not only a good reconstruction of the folk conception, but is also psychologically realistic, we will have shown that, in one area at least, folk-psychology has been vindicated by our advancing understanding of the mind/brain.

7.6. Support from the Neurosciences.

In this section I shall suggest that my hierarchical theory of moods and traits mirrors the neural structure of the brain. Thus, while it may be possible to offer behaviourally equivalent models of the mood system that use only a single functional level, mine is more likely to be psychologically realistic. I take this to be a matter of some importance, since, although philosophical psychology may be interested in a maximally general characterisation of the mental, scientific psychology aims at descriptions of the minds of human beings and other particular organisms. It is when we are in pursuit of this goal that we rule out, for example, theories of the mind that call for more computational steps than a biological system could carry out in real time. We bring our knowledge of the physical realisation of the system to bear to constrain our functional theory of it.

Strong and Weak Realisation

If I am correct, my hierarchical model is "realised" in the brain in a stronger sense than that in which single level models are "realised". A functional model is weakly realised in a particular system just if it gives an input/output adequate description of that system. Thus, any adequate functional description of a system is realised by
that system. A functional model is strongly realised in a particular system at some lower level $L_n$ just if there is an independently motivated taxonomy of that system, at level $L_n$, which maps the structure of the functional model. It is necessary that the taxonomy of $L_n$ be independently motivated, as it will always be possible to gerrymander a taxonomy of $L_n$ that allows such a mapping to be constructed. An independently motivated taxonomy is one that is constructed without the guidance of the higher level model that we are seeking to map. Instead, it should emerge naturally from a description of the system at $L_n$. In the case at hand, I take a neurological account of the brain to be independently motivated. If such an account gives us a taxonomy which is isomorphic to the internal structure of a higher level functional model, then that functional model is strongly realised at the neural level.

**Neuroscientific Evidence**

There are a number of pieces of research which suggest that moods are chemical states of the central nervous system which affect the probability that a certain transition will take place, and that traits are intra-individual differences in some of the same systems. In reading what follows we should bear in mind the evidence we considered in chapter three for the localisation of the affect-programs in the limbic system, and especially the hypothalamus. These results connect promisingly to work on the neurological basis of mood and trait phenomena.

Although I take the co-incidence of these two groups of results to be a point in favour of our theory of moods, I don't want to suggest that our theory of moods is restricted to the sort of emotional phenomena to which the affect-program account is suited. The affect-program states are relatively well understood in comparison to other affective phenomena. This makes it possible to relate them to other areas of neural research, in a way which, though speculative, is reasonably well grounded. The fact that we can't similarly relate the other emotional phenomena reflects our comparative ignorance of their neural basis, rather than any lack of connection.
Chapter Seven: Moods and Character Traits

The classic work on the neurological basis of mood and trait phenomena has concentrated on depression, since this is a state of particular clinical interest. Much of this research has been a result of the famous "catecholamine hypothesis" (Schildkraut 1965, 1970, 1974). The catecholamine hypothesis concerns the neurotransmitters norepinephrine, dopamine and serotonin. The connection with the work on emotion reviewed in chapter three is that neurons in the limbic system are specialised to these transmitters, concentrations of norepinephrine being at their highest in hypothalamus. Under certain circumstances, drugs which decrease levels of these neurotransmitters, thus reducing the facility of synaptic transmission in the neurons that utilise them, appear to induce depression. Others, including cocaine, which inhibit the removal of these transmitters after their release, appear to produce euphoria, or to relieve depression. The catecholamine hypothesis suggests that depression is produced by neurotransmitter deficit, and more tentatively, that mania is produced by oversupply. Although the hypothesis has lead to a large body of interesting results, it is now generally recognised to be inadequate to the data available (Baldessarini 1975). Even workers most sceptical of the hypothesis itself, however, maintain that changes involving these neurotransmitters play a role in mood phenomena.

A still more direct connection between work on the affect-programs and these neurotransmitters is to be found in J Panksepp (1982). As we discussed in chapter three, section six, Panksepp has tried to map the circuits responsible for affect-program responses in the limbic system. He proposes that "Biogenic amines that are widely dispersed throughout the brain may provide nonspecific excitatory and inhibitory control over emotive command circuits." Panksepp is thinking in particular of serotonin and norepinephrine. Increased availability of these transmitters may, he thinks, have a respectively inhibitory and excitatory effect on
the whole affect-program system. Other neurotransmitters, acetylcholine and dopamine, are linked to particular affect-programs in Panksepp's list.

Another area of research on depression is work on neural electrolytes. Sodium ions play a central role in neural functioning, and depression is associated with the importation of sodium into cells throughout the body. Lithium, an ion importantly similar to sodium, has long been recognised as an effective therapeutic tool in treating affective disorders. Moods are thus once again linked to broadly diffused effects on synaptic transmission.

Finally, there has been a body of work on the endocrine system. Suggestive results have been obtained by E.J. Sachar (1970, 1975). He replicates the well known result that certain groups of depressives hypersecrete cortisol, and suggests that this may be linked to the neurotransmitter effects just cited. Neural tracts dependent on those transmitters are known to be involved in cortisol production, and depletion of the transmitter norepinephrine in animals can stimulate the endocrine pathway involved in cortisol production. These links to the endocrine system are particularly interesting because of the importance of endocrine gland malfunction in long term personality disorders.

Such results as these make it plausible to suggest that moods are neurochemical states, which act to modify the activity of broad areas of the central nervous system. If this is how the neuroscientific picture turns out, then their action would be precisely to modify the probability of transitions between a given input, internal states and output. Very simply, suppose that under certain conditions, the presentation of a spider has a high probability of starting the affect program for the fear response. The effect of depression would be to inhibit the depolarisation of neurons in the hypothalamic pathway subserving that response, and thus to alter the

---

1 While we have expressed reservations in chapter three about Panksepp's typology of the affect programs, these criticisms would not bear on the nonspecific role for serotonin and norepinephrine.
transition probability between the presentation, and the fear response. This sort of story, if true, would fit in well both with the theory, and with the account of the folk use of mood vocabulary given in section five.

When we turn from moods to traits we find a connected body of research. Studies of the endocrine system, known to be an important factor in personality disorders, suggest that long term individual differences in the endocrine response may underlie long term psychological traits. K.H Tennes and J.W Mason (1982) summarise a large number of such studies. To take one example, it has long been known that hormones such as cortisol and growth hormone show elevated production under stress. These elevations, however, are imposed on large differences in base levels across individuals. These base differences seem to correlate with individual psychological differences, such as the deployment of certain characteristic defence mechanisms in stress situations. Variations in endocrine response are known to emerge very early in life, but there is as yet no clear indication of the relative importance of genetic and environmental factors.

If the picture of the mood/trait system suggested above is correct, then we can see why moods and traits function like higher order states. Emotions, or at least, those dealt with by the affect-program account, are neural states within a modular system. These states are triggered by certain inputs, and give rise to characteristic patterns of physiological change and short term behaviour. A mood state is a neurochemical condition which modifies the propensities of one neural event to bring about another. It thus alters the functional description realised by the affect-program system, its inputs and its outputs. Its effects may be diffuse, affecting the whole system, or specific, affecting the various emotional responses differentially. In addition to its effects on the affect-program system in the limbic brain, the same chemical condition may affect higher areas of the brain, and thus affect emotional phenomena involving higher cognition, and higher cognitive phenomena generally. It may also have direct physiological and behavioural effects, perhaps by direct
influence on output systems. Traits, I have speculated, may turn out to be longer term states of the systems responsible for these neurochemical conditions. Supposing that some traits are background levels of a given chemical, against which mood fluctuations occur, it is possible to see how a trait might have direct emotional and behavioural effects, while also affecting our propensity to exhibit certain moods. Given this physiological picture, the sort of hierarchical model we have suggested at the functional level will directly mirror the causal structure of the brain.

7.7. Conclusions

I set out to support the proposal that moods and traits are higher order functional states. On this view, the system of moods and traits serves to explain the different functional descriptions realised by different persons, and by persons at different times. We have seen that folk discourse does indeed use mood and trait terms for this purpose. We have also outlined some data on the possible physical realisation of the mood/trait system. We have suggested that changes in a person's moods may correspond to changes in the presence of certain neurally active chemicals in certain regions of the brain, and that some of the differences between persons referred to as personality traits may be realised by long term individual differences in the systems responsible for the production of some of those neurochemicals. Together, these two bodies of evidence lend support to our model of moods. The model captures some features of the folk conception, which entitles us to call it a model of moods and traits, and it is sufficiently in harmony with what we know about the neural underpinnings of mood and trait phenomenon to suggest that it may be strongly realised in the brain. If the neuroscientific picture I have described is correct, then the idea that certain tracts of the brain change their functional characteristics across time mirrors the proposal that mood changes are
changes in the functional description realised by the person. The idea that some inter-individual, and longer term intra-individual, differences are due to biases of the mood system mirrors our model of character traits as a still higher functional level in a larger model of persons.

There is one aspect of the model in which it may appear to depart from the folk conception quite drastically. In the folk conception, people may be in several moods at once, and have many traits of personality. But if a person's mood is the overall state of his/her neurochemistry, one might think that he/she would only be in one mood at a time. Similarly, one might think that the long term condition of the endocrine systems would yield a single personality condition, rather than a series of independent traits. However, I don't think either of these suppositions have to be correct. In the case of moods, different chemicals affect different areas of the brain, and different aspects of functioning. Electrolytic phenomena may affect all neural transmission, but a depletion of norepinephrine would differentially affect different tracts of neurons. Whether there would be any correspondence between our folk categories of moods and various independent variables in brain chemistry is an open question. In the case of traits, there are many systems which contribute to the long term state of the brain's chemistry. These might be differentially affected by genetic factors, or by the early environment. It is another open question whether there is any detailed correspondence between this structure and our vocabulary of traits. Admittedly, introducing such matters at the functional level might call for a more complicated model than that introduced here.
Appendix I - Experimental Techniques

(Chapter 3)

In chapter three I presented the results of some cross-cultural judgement tests, and of a component analysis of spontaneous facial expression. I omitted many details of these experiments to aid the exposition, and gave only the salient results. Those who find these results implausible may wish for more detail of the way these experiments were conducted. In this appendix I give details of the most conclusive judgement test, and the component analysis.

LI Ekman & Friesen (1971) - Cross-Cultural Judgement Test

As noted in chapter three, Ekman and Friesen used a judgement test devised for use with children. Each observer was presented with three photographs of a western face, chosen by the experimenters to represent three distinct emotions. The observer was then told a story designed to describe a suitable situation, in his own culture, for the occurrence of one of the emotions pictured. The observer then pointed to the face he or she considered most appropriate to the story. This technique has the advantage of avoiding the translation of emotion terms from one language to another, a procedure fraught with difficulties.

To avoid the possibility that the observers shared the experimenter's cultural conventions about emotional expression, they were chosen from the most culturally and visually isolated group available. The observers were members of a tribe of the Fore language group in New Guinea. The individuals chosen had seen no movies or magazines, were unable to speak English or Pidgin and had never lived with or worked for westerners. A set of 40 photographs of 24 different persons were used, and they were shown to 189 adult and 130 child observers.
The results are summarised in table 1. They are clearly very supportive of the hypothesis that visually isolated New Guinean subjects can accurately infer the emotional state of westerners from their facial displays. The only problem encountered was that the New Guineans appeared to have difficulty in distinguishing fear from surprise.

<table>
<thead>
<tr>
<th>Emotion Category Described in Story.</th>
<th>Adults</th>
<th>Children</th>
</tr>
</thead>
<tbody>
<tr>
<td>Happiness</td>
<td>92</td>
<td>92</td>
</tr>
<tr>
<td>Sadness</td>
<td>79</td>
<td>81</td>
</tr>
<tr>
<td>Anger</td>
<td>84</td>
<td>90</td>
</tr>
<tr>
<td>Disgust</td>
<td>81</td>
<td>85</td>
</tr>
<tr>
<td>Surprise</td>
<td>68</td>
<td>98</td>
</tr>
<tr>
<td>Fear...</td>
<td>80</td>
<td>93</td>
</tr>
<tr>
<td>.....from Anger/Disgust/Sadness</td>
<td>43</td>
<td>100</td>
</tr>
<tr>
<td>Total No of Observers</td>
<td>189</td>
<td>130</td>
</tr>
</tbody>
</table>

The results of the Ekman and Friesen experiment have been reinforced in two ways. First, the experiment was repeated by a different team with a still more isolated New Guinean culture, and similar results were obtained. Reports of this data can be found in Ekman (1972). Secondly, the Fore speakers were asked to act the role of the people described in the emotion eliciting stories, and unedited videotapes of their behaviour were shown to university students in America. The students were asked to guess which emotions the Fore speakers were acting out. Results are shown in table 2. Once again, a high degree of cross-cultural recognition was achieved. The New Guineans failure to distinguish fear from surprise, however, shows up in these results as well.
Appendix I: Experimental Techniques

Table I.II  % Correct Judgements by U.S Observers of New Guinean's Intended Emotions. (From Ekman 1972)

<table>
<thead>
<tr>
<th>Emotion</th>
<th>% Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Happiness</td>
<td>73</td>
</tr>
<tr>
<td>Anger</td>
<td>51</td>
</tr>
<tr>
<td>Disgust/Contempt</td>
<td>46</td>
</tr>
<tr>
<td>Sadness</td>
<td>68</td>
</tr>
<tr>
<td>Surprise</td>
<td>27</td>
</tr>
<tr>
<td>Fear</td>
<td>18</td>
</tr>
</tbody>
</table>

II Ekman, Friesen & Malmstrom (1970) - Component Analysis

In chapter three I described a component analysis of the facial expressions of Japanese and American university students. The expressions were measured by the FAST method (Facial Affect Scoring Technique). FAST consists of an "atlas" of prototype photographs. Films or photographs of the faces to be judged in an experiment are compared to these prototypes to determine which emotion they express. The background to FAST is a theory, a components hypothesis, about which facial muscle movements express which emotions. The emotions involved are Ekman's usual list - happiness, anger, disgust, sadness, surprise and fear. For the purposes of FAST the face is divided horizontally into three areas, brows/forehead, eyes/lids and lower face. The atlas contains eight shots of brows/forehead, seventeen of eyes/lids and 45 of lower face.

To assess a picture of a face, it is divided into the three areas. With films, Ekman's preferred technique, each discernible facial movement is isolated, and its duration assessed. A group of judges then class the area or areas in which the movement occurred with the FAST prototype of that area which they most closely resemble. Formulas are applied to these judgements to determine which emotion, if any, is expressed by the movement. When a film has been analysed, the results show both the frequency and the duration of each emotion as displayed in each area.
of the face. Further formulae combine these results for areas into results for the total face.

The important fact about FAST is that there is no call for any judge to assess which emotion is being expressed. All the judges assess is the visual similarity of a picture to some other pictures of that area. Relating this data to judgements of emotion is left to the theory, which is embodied in the choice of prototypes and in the formulae used in combining the judgements. FAST allows us to put a theory about facial movements into a form where it can be consistently and conveniently applied to any body of filmed data.

In the experiment described in chapter three, Ekman et al attempted to show that the same facial behaviour is used to express the same emotions in American & Japanese subjects. Averill, Lazarus and Opton (1969) showed that Japanese and American subjects reported similar emotions on watching stress inducing motion picture films. FAST was used to show that they used similar facial behaviour to express these emotions. Twenty five American and twenty five Japanese students were shown the films while alone in a room. They were aware that their skin conductance and heart rate were measured, but unaware that they were being videotaped. The results are shown in table 3. It can be seen that they show a large degree of similarity between the two group's facial behaviours.

At the time of this experiment, Ekman et al did not feel that sufficient validation studies had been conducted on their version of FAST to allow them to be certain that the facial behaviour classified as emotion x by FAST really was emotion x. However, the design of this particular experiment enabled them to by-pass this problem. Since it was already known that the Japanese and Americans gave self-reports of similar emotions when confronted with this stimulus, the fact that they displayed similar facial behaviour confirms the hypothesis that they use similar facial behaviour for each emotion.
### Table I. III Frequencies of Emotional Facial Behaviour of American & Japanese University Students under Stress condition. (From Ekman 1972)

<table>
<thead>
<tr>
<th></th>
<th>Total Face</th>
<th>Brows/FH</th>
<th>Eyes/lids</th>
<th>Lower face</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>US</td>
<td>J</td>
<td>US</td>
<td>J</td>
</tr>
<tr>
<td>Surprise</td>
<td>76</td>
<td>50</td>
<td>53</td>
<td>24</td>
</tr>
<tr>
<td>Disgust</td>
<td>61</td>
<td>48</td>
<td>16</td>
<td>13</td>
</tr>
<tr>
<td>Sadness</td>
<td>59</td>
<td>126</td>
<td>29</td>
<td>39</td>
</tr>
<tr>
<td>Anger</td>
<td>29</td>
<td>28</td>
<td>18</td>
<td>28</td>
</tr>
<tr>
<td>Happiness</td>
<td>8</td>
<td>14</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Fear</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Non-Specific Facial Behaviour</td>
<td>29</td>
<td>24</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

**Rank Order Correlation**
- 0.88 p<0.01
- 0.86 p<0.05
- 0.95 p<0.05
- 0.96 p<0.01

As mentioned in chapter three, a further result of the experiment was that introducing an interviewer into the room caused the Japanese to mask their emotional responses to the films, and to smile politely for much of the session. Traces of their previous responses were still detectable by micro-momentary analysis of the videotapes. This example of a "display rule", a cultural norm requiring suppression of emotion in certain social situations, may help to explain the widespread belief in the emotional "inscrutability" of the Japanese.
Appendix II - The Sociobiological Option

(Chapter Five)

II. Sterelny's Argument

In chapter five I discussed certain possible uses of sociobiology in emotion theory. I showed that if some emotions are sets of propositional attitudes, explanation is needed of why those particular states occur, and not the vast multiplicity of other sets that would seem equally eligible. I suggested that these cognitive emotions may occur cross-culturally, and that if they do, sociobiology might provide explanations of their occurrence. The writings of some sociobiologists contain attempts at such explanations. There is however, a general argument, suggested by Kim Sterelny (1987), that would, if successful, doom all such attempts to failure. This argument is of particular interest, because it suggests that there is a principled difference between evolutionary explanations of modular systems, such as the affect-program responses, and attempts to give evolutionary explanations of cognitive emotions.

Sterelny argues that there cannot be selection for most human behaviours, because there is no single, dedicated mechanism responsible for each of those behaviours. He pictures selection for behaviours which are mediated by dedicated modules as driven by the selective advantage which accrues to genotypes which build neural structures that serve wholly or predominantly to bring about those behaviour. He anticipates evolutionary explanations of the various modular sub-systems, such as the sensory input analysers, and perhaps the language processor(s).

Much behaviour, however, is not mediated by any dedicated mechanism. Most intelligent human behaviour seems to be mediated by a general purpose intelligence, usually represented as a system which combines beliefs and desires in virtue of their contents, so as to rationalise action. I shall conduct my discussion as if this
traditional picture is correct, but I hope that what I say will be of wider relevance. Certainly, the insight Sterelny has does not depend on the belief/desire model. It rests on the more general idea that many kinds of intelligent behaviour are mediated by one set of mechanisms, whatever those mechanisms may be. Where there is no specific behaviour producing mechanism for the genes to build, argues Sterelny, that behaviour cannot be selected for.

Sterelny's insight is that if one set of mechanisms subserves so many functions the alterations which improve its performance of one function will interfere with its performance of other, equally important functions. Let us consider this problem using the traditional picture of the mind as combining beliefs and desires in virtue of their content to produce action. On this picture, all that is genetically coded, and therefore all that can be selected for, are the set-up parameters of the cognitive system. The system deals with all content domains, and any innate information, principles of rationality, or principles of belief formation which it embodies (the set-up parameters) must apply to all content domains. Any change to the innate structure of the system will therefore have all sorts of global effects, and is vanishingly unlikely to promote any small set of useful behaviours without messing things up elsewhere. Consider an attempt to alter the system so as to make an animal more sexually aggressive to its conspecifics. In an intelligent animal, simply being more aggressive isn't the sort of thing that would be of selective advantage. "Being aggressive to other males who aren't dominant, and when there's something to be fought over" might be, but that's not the sort of thing that could be selected for. Selection at the level of set-up parameters just isn't fine-grained enough. So there can't be selection for behaviours which are subserved by a multi-purpose cognitive system.

Another way to put Sterelny's point is that the mechanisms that mediate intelligent behaviour are likely to be horizontal, rather than vertical. Rather than being content domain specific, like some postulated language processing modules, or the affect-
programs, they will be specific to some sort of function. There are, on this suggestion, mechanisms for short term memory, imaging and image transformation, motor control, etc, rather than modules for mathematics, feeding, aggression, and the like. Thus, the same set of mechanisms would be involved in feeding behaviour, aggression towards conspecifics, and literary composition.

II.II Reply to Sterelny.

Sterelny's argument contains far too many quantitative assumptions to be the sort of argument that can be assessed a priori. We don't know how polymorphic a trait our central cognitive processes are, and so have little basis for estimating the "fine-grainedness" with which genetic variation could occur. The extent of our ignorance here is quite remarkable. In the case of simple organisms like insects it is clear that genetic control over the structure of the nervous system is pretty tight. Yet even in this simple case there is as yet no good account of how the neuron produced at a particular cell division comes to be connected in the necessary way. Except in the very simplest organisms it does not seem that the neuron produced at any particular division is destined to play a particular role. The ability to produce a normally functioning central nervous system seems quite mysteriously robust under certain sorts of developmental abnormality. In human beings, the genes seem able to produce a functionally equivalent structure in a foetus born hydroencephalitic, with a quarter of the usual brain weight plastered in a thin layer around the inside of the skull.

Another way to show how dependent arguments like Sterelny's are on quantitative assumptions is to show that an argument against the possibility of certain types of morphological evolution can be constructed, which parallels the argument under consideration. After all, the genes only directly determine the production of proteins. For any genetic change that might occur, is it not vanishingly unlikely that
the small group of desired changes will result without messing things up elsewhere? The problem with all such arguments is that they assume that a vast array of variables have particular values.

Let's consider the upright stance of humans an example of the morphological version of the argument. It is pretty uncontroversial that our upright stance is adaptive. Many accounts of how it is adaptive are on offer. It lets us see further, frees our hands, etc. The stance example is interestingly parallel to the sort of cognitive changes we are discussing, in that it is a alteration at the margin to a previously existing, and highly complex, system. The upright mammal and its four-footed predecessor still retain a vast morphological overlap. The skeletal/muscular mechanisms used to achieve the stance are adaptations of those used by the other primates in their very different stances. The perceptual and neural mechanisms, such as those of the inner ear, are also adaptations of existing mechanisms for balance, co-ordination, etc. Presumably, some small genetic change was once such as to affect enough of these systems at the same time to create, if not the full upright stance we have now, at least enough of a stance to confer significant advantages. The same change will have had many undesirable consequences, such as the narrowing of the birth canal, and the dangerous compression of the discs in the lower back, but overall, the effect of the initial adaptation managed to be advantageous. It is all too easy to see how, say, a creationist, could represent as vanishingly unlikely the notion that a genetic change could be such as to cause all the appropriate changes, and no compensating disadvantageous changes.

Let us now turn to higher cognitive processes. Take some sophisticated cognitive system. It consists of a whole series of horizontal modules surrounded by input modules, and certain vertical faculties, such as the language processor, and the affect-program system. A small genetic change produces a protein which affects various of the modules in lots of different ways, and has the effect, among others,
of creating, say, an alteration in the principles of belief formation. What is to stop this bearing differentially on the class of internal states which realise some fitness relevant class of contents? Suppose it were to create a previously non-existent moralistic thought process, and thus contribute to fitness through reciprocal altruism. The fitness increase from this might outweigh any other, less fortunate effects the change has on the workings of the cognitive system. We can be told that a change which has just these effects is vanishingly unlikely. But why is this any more convincing than a creationist attack on the selective explanation of the upright stance?

Sterelny has suggested in conversation that the numbers may be on his side, while not on the creationists, because the morphological change created by an alteration to the genotype is fixed once and for all during development, whereas the cognitive change must continue to interact holistically with the rest of the cognitive system whenever the animal thinks. I fail to see that there is any real disanalogy here. In both cases the genetic change interacts with the rest of the genotype during development to produce a given phenotypic change(s). For the rest of the animals life, this change or changes interacts with the rest of the animal's phenotypic features. In the cognitive case this happens whenever it thinks, and in the morphological case whenever it acts.

II.III A Counterexample to Sterelny

I have argued that Sterelny's argument against the evolutionary explanation of intelligent behaviour mediated by general purpose cognitive mechanisms depends upon a range of quantitative assumptions, and that there is no reason to suppose that these hold. I now want to give one example where it seems clear that they can't hold, as intelligent behaviour does seem to be genetically determined. The example I have in mind is the occurrence of characteristic species-specific patterns of social and sexual organisation in primates. The reason for thinking that these patterns of
organisation are under genetic control is that they continue to occur in considerably altered environments (for example in captivity) Gorillas simply don't exhibit the extended troop behaviour and unstable leadership of chimpanzees however much you alter their environment. Even when social behaviour is altered by changes in the environment, it is in ways that seem best explained as the result of the same inner forces acting in different circumstances, as when a captive chimpanzee troop with no mature males retains its characteristic structure, but with a dominant female as leader. But although the social behaviour of these primates appears to be strongly influenced by their genes, it also appears to be mediated by their most intelligent cognitive processes, the same processes that are responsible for their remarkable achievements of tool-making, deception, alliance-forming, and so on. In fact, some authors (Humphries 1976) have suggested that the main evolutionary advantage accruing to primate intelligence is enhanced ability to deal with complex social interactions. In the example of primate social behaviour we seem to have a prime counterexample to Sterelny. A flexible, intelligent cognitive system is deployed to organise the social interactions of a species, but the set-up parameters of that cognitive system are such as to impose a species-specific pattern of social organisation.
Appendix III : A Theory of Sensations

(Chapter Six)

III.1 Introduction

In chapter six, section one, I said that I would append an account of sensations that sidesteps the traditional, behaviouristic antipathy characteristic of philosophical cognitivists and constructionists. Philosophers like Armon-Jones still seem to consider it important to deny that an emotion could be recognised on the basis of a quality of sensation associated with it, or that the identity of an emotion could depend upon such a quality. The modern psychoevolutionary approach to emotions, on the other hand, takes an emotion to be a pretheoretically vague collection of phenomena, and makes no essentialist claims. On this approach, if emotions lead to sensations, they may play all sorts of important roles. In chapter six, section three, I discussed the work of the neuroscientist P.D MacLean, and the psychologists C. Izard and S. Tomkins, all of whom think that we can identify emotions by the qualities of sensation associated with them. Tomkins and Izard also suggest that the sensation associated with an emotion may be acceptable or unacceptable in virtue of its subjective quality, and thus act to reinforce behaviours positively and negatively correlated with the emotion. I argued that philosophy should seek to provide a coherent interpretation of talk about qualities of sensation which will allow qualities to play the role they seem to play in some psychological theories.

In this appendix I present a physicalistically acceptable theory of sensations which tries to make as much room as possible for what those untouched by philosophical worries want to say about subjective experience. I talk mainly about perceptual sensations, because the particular neuropsychological theory which I use to make my discussion more concrete was introduced with perceptual sensations in mind.
But the general, philosophical points are not tied to this model, and should be equally applicable to interoceptive sensations, including those associated with emotions.

The idea that phenomenal experiences should be identified with precise physical states of the nervous system, rather than states at some higher functional level, is a commonplace. The suggestion has been made in many places, perhaps most notably by David Lewis in "Mad Pain & Martian Pain" (1980), and is now one of the standard functionalist responses to the "problem of qualia". But these well-known proposals all use the traditional model of levels of nature. They distinguish the functional level, or levels, from the level of physical realisation, a level whose terms are supposed to denote items of ultimate physical stuff, incapable of multiple realisation at any lower level. They presume that a neuropsychological theory would identify qualia with states of this ground level, physical stuff.

Here, however, I take a Lycanesque approach to levels of nature (Lycan 1987). The neural level is just one of the many functional levels at which reality can be described. Neural descriptions are multiply realisable. The same neural network may be instantiated by two systems, one of which uses chemical and the other electrical synapses. Each of these systems may, in its turn, be capable of realisation in a number of ways, for example by lower level systems using different neural electrolytes. My account will be neutral between a strong Lycanesque approach, which conceives nature as "functional all the way down", and a weak Lycanesque approach, which merely supposes that any non-functional level is below the level at which the neurosciences are conducted. The difference between these two positions should not affect the matters addressed here.

The Lycanesque approach makes the move from a traditional, functionalist theory of qualia to a neuropsychological theory far more interesting. It becomes apparent that two different positions are confounded in the claim that the correct theory of
qualia will occur at the neural level, or "the level of physical realisation". The first of these positions is that identical qualia are guaranteed where a system is identical to us at every possible level of description, however low. The second is that identical qualia are guaranteed where a system is identical to us down to the level at which a certain neuropsychological account is given. It is this second position that I shall defend. This position is open to reformulated versions of some of the traditional qualia-based objections to functionalism, but the added resources of the neuropsychological level allow us to counter these objections. The neuropsychological account will explain the nature of phenomenal experience in such detail that it will become increasingly hard to claim that something has been omitted.

In section two I outline the sort of neuropsychological account I have in mind and what it can explain about phenomenal experience. In section three I argue that the theory gives us a basis for attributing phenomenal experiences to other systems similarly constituted to us at and below the neuropsychological level. In section four I argue that any system which shares our neuropsychology, however it is realised at lower levels, has phenomenal experiences similar to our own. In section five I consider the objection that these phenomenal experiences would be causally inefficacious epiphenomena. I show that there is a paradox inherent in the view that phenomenal experiences are epiphenomenal. In section six I avoid causal inefficacy by arguing that phenomenal properties are constituted by neuropsychological properties, and do not amount to anything over and above those properties. As well as solving the problem of causal inefficacy, this move strengthens the case for saying that any system of which the neuropsychological theory is true has phenomenal experiences, no matter how its neuropsychology is realised at lower levels. Finally, I confront the recalcitrant intuitions of some hard line defenders of the irreducibility of qualia, and consider the similarities and differences between my theory and some traditional functionalist attempts to deal
Appendix III: A Theory of Sensations

with phenomenal experience. I then assess to what extent my theory is able to reconstruct the claims of MacLean, Izard and Tomkins.

III.II Neuropsychology & Sensation

There is not yet a definitive theory of sensation at the neuropsychological level. Pribram (1971) has suggested that sensations may be complex structures of slow potentials in certain types of neural network. Churchland (1986) has proposed the neural hyperspace hypothesis, which we will discuss in some detail. Neither proposal could really be said to be more than informed speculation, but informed speculation will be enough to fulfil one very important function. Using these speculations we can explain how a purely neurological account of sensation might one day account for much of the phenomena of introspection.

Churchland's speculations take as their starting point the fact that all sensory modalities produce sensations by integrating information from a number of receptor types. Thus, a given colour sensation results from integrating the response of the three types of cone in some portion of the retina, and the taste of something from integrating the responses of the four types of receptor in the tongue. The end state produced by such an integration is nicely represented as a position in a multidimensional space. Such spaces might be produced neurally in a number of ways. Very simply, suppose we have a set of n channels of input from a sensory modality, where n is the requisite number of dimensions. Let the spiking frequencies in the fibres of each channel set the value of a co-ordinate. A particular sensation can then be identified with a position in the resulting hyperspace.

Churchland's proposal passes a few simple empirical tests. It explains, for example, why we have so many more taste than colour sensations (it predicts an order of magnitude more), and how it is possible for a dog to have such an extraordinary ability to distinguish smells - the hyperspace notion allows a fairly simple neural array to represent a vast number of distinct positions. We can also
represent the three common forms of human colour-blindness, using the three possible ways in which the colour space could collapse into two dimensions by the loss of one input channel (axis).

But the most important feature of this proposal for our purposes is its ability to give satisfying explanations of various introspective phenomena. To take the simplest case, similarities and contrasts amongst sensations are explained as proximity or otherwise in the hyperspace. We can explain, for example, why the visual sensations produced by the colour spectrum seem to move along a gradient of similarity. Of course we already have an evolutionary story about the desirability of interpreting similar kinds of reflected light as similar perceived qualities. The neural story explains how this trick is turned. It tells us what it is about crimson and violet that makes them seem so phenomenologically dissimilar - the relatively large distance between them in the neural space. In general, our neuropsychological theory will give explanations of relationships between phenomenal properties which we have previously been forced to treat as brute facts of similarity and difference.

Consider next the phenomenon of inter-modal similarity judgements. Much of our aesthetic activity involves such judgements. Baudelaire's sonnet "Correspondences" famously puzzles over this fact. The sounds of trumpets are like silver, other sounds are systematically described as dark or light. What non-mystical story can we possibly tell about such "similarities"? More mundanely, we have a vocabulary of pain sensations which classes them as warm, bright, dull, etc. We also compare certain sexual sensations to pain sensations, or to sensations of heat. Some of these correspondences can be linked to associative processes, although it's not clear that this is a full explanation1. Others cannot even be so linked.

---

1 Once again, we have an etiological explanation of why the sensations should be similar, but no story about what their similarity consists in.
We can give a good account of these phenomena using the hyperspace story. It is a central feature of the sort of theory of neural computation in which the notion of neural spaces is originally introduced, that various spaces map onto one another in ways that facilitate certain mental processes. In the case of sensory spaces this sort of mapping may occur as part of the process of producing a unitary image of the environment. Imagine the various sensory spaces being mapped onto one another in a way, or ways, that reflects previously experienced contiguities amongst sensations. Part of this mapping (or some of these mappings) might be produced during the life of the organism, part (or some) might be innate. Such mappings will establish a projection of one space onto another, presumably in a way designed to preserve a pattern of links between various points in the two networks. This would map many points onto one another as a mere artefact of the projection, thus causing us to have a whole range of similarity judgements that reflect, apparently, nothing but the nature of the sensations themselves, as well as judgements that reflect experienced associations. The phenomenological similarity of the sensations is explained by their respective spaces being mapped onto one another in this way. Once again, what was phenomenologically brute is explained and validated at the neuropsychological level.

As well as explaining intra-modal and inter-modal similarity judgements, the hyperspace story will allow us to reconstruct the often, but obscurely, referred to idea that the "unanalysable" qualia may have structure. A taste may be felt to have bitterness as an especially strong element, and this can be seen to arise from it's being very far towards one end of the bitterness axis in the four dimensional taste-space. Similar stories can be told for the other modalities.

It does not matter that these putative explanations are based on speculations. My aim is not to argue for the truth of the hyperspace theory, but to show that the structure and relations of qualities of sensation might one day be accounted for in detail by a neuropsychological theory. The purpose of our speculations is to make
it easier to get a grip on this claim. Such a theory will explain what it is about red that makes it so unlike violet, and what is bright about the sound of a trumpet. The phenomenologically discernible structure of a taste, or a shade of colour will also be explicated in detail at the neural level.

It has been generally assumed that the arrival of such an intuitively satisfying neuropsychological account would mark a bad day for the qualia freaks. But this is a mistake. I shall argue that, far from showing that there are no qualia, the account would show that there are qualia, and show precisely what qualia are.

III.III The Conservative Inference

Suppose the theory we’ve been describing turns out to be correct. I know that similar processes to those described in the theory are going on in my own brain. But I also know that when I make the reports that people make when they have these brain events, I am reporting my phenomenal experiences. When I report that saccharin tastes different from sugar, I am at least partly reporting differences in the subjective quality of my experience. I may well see that the difference lies in a greater element of bitterness in the taste of saccharin, but that is not to deny that the two differ in subjective quality, it is merely to specify the difference.

So the theory will reveal to me which states of my nervous system support my phenomenal experiences. Knowing that others share these states, I am forced to infer that they share my phenomenal experiences. But the inference that others share my phenomenal experiences can be made with two degrees of generality. First, I might make the conservative inference that all normal humans have similar phenomenal experiences. Secondly, and more radically, I might infer that any system similarly constituted at the neuropsychological level has similar phenomenal experiences. These two inferences correspond to the two interpretations,
distinguished in section one, of the claim that the proper theory of qualia identifies them with specific neural states.

The first, more conservative inference is less open to objection, because our neuroscientist can argue that other humans have the same qualia in virtue of being identical at every level of description which could possibly be relevant. The cells of different brains will, of course, consist of different spatial arrangements of molecules, but the idea that this could affect phenomenal experience will attract few adherents. Other than this, it is quite legitimate to abstract away from individual differences in developing a general theory of the mind. A theory of visual processing is not refuted by diseases and abnormalities of the visual nervous system. So we can argue that the conservative inference is legitimated by a plausible principle of supervenience. Two physical systems identical at every level of description at or below the neural shouldn't display different properties above that level.

While I believe that the conservative inference is a sound one, it is not a very interesting one. Few people are tempted by a sceptical line about the introspective life of normal humans. The real problems with phenomenal experience emerge when we want to attribute a fully fledged mental life to systems that are identical to humans at some level, but are differently realised below that level. This brings us to the more daring inference that was opened up by the success of the neurosciences. We could infer that any system of which the neuropsychological theory is true has the same phenomenology as a normal human being. In the next section, I shall consider how accepting this inference would affect attributions of phenomenology, and defend its validity.

III.IV The Radical Inference

Suppose we discover that people who are brought up on the moon in the settlements of the next century have a different isotope of some chemical playing a
role in their neurochemistry. This difference causes no changes in the firing probabilities of their neurons, or in any more diffuse electrical effects, so a neuroscientist would probably consider it an interesting irrelevancy. Nevertheless, it does slightly reduce the strength of the claim that the principle of supervenience guarantees that they have normal phenomenal experiences. It gives a tiny foothold to someone who would deny this.

Now suppose that we are introduced to an android model of Baudelaire in a twenty-first century Madam Tussaud's. Metal Baudelaire passes the Turing test, and is as fascinating and perverse as you could wish. However, the way in which he processes sensory information is quite different from the way in which we do. No inter-modal similarity judgements emerge from the structure of his sensory system. To get him to write symbolist poetry for the tourists, it has been necessary to give him an ancillary structure which lists a series of basic metaphoric connections culled from his predecessors opus, and recursively derives striking sensory metaphors for the subjects he wishes to write about.

In this case, the claim that the android's sensations might not have the same phenomenal qualities as ours seems less of a sceptical cavil. There is no inner state of Metal Baudelaire that plays the neuropsychological role of the sensation of scarlet, or the sound of middle C played on a trumpet. His dispositions to say that sensations are similar have nothing to do with the nature of the inner states which constitute for him the registration of the presence of something red, or trumpet-sounding. Once we understand Metal Baudelaire we see that there is nothing about his "sensing high notes on a trumpet" state that would give it a perceived similarity to his "sensing something silvery" state. But in that case, we have reason to doubt the claim that his phenomenal experience of those states has the intrinsic similarity that ours has. As we have already noted, claims of brute phenomenological similarity stand in need of explanation and validation in terms of the inner states which they claim to report.
But if Metal Baudelaire's inner states lack the relational features of the human ones, it becomes hard to see how he could have the same phenomenal experiences. A "burning pain" that is not intrinsically more similar to a sensation of heat than to, say, the taste of sugar, wouldn't be the burning pain that figures in my phenomenology.

There are two lessons which seem to emerge from these stories. First, the higher the functional level below which identity fails, the more grip there seems to be for sceptical doubts about the phenomenal experiences of that system. Secondly, a good level at which to stop treating these doubts as sceptical cavils, and start treating them as serious worries, is the neuropsychological level, where we get our explanations of the structural and relational features of sensations.

This level is significant because these structural and relational features seem, phenomenologically, to be intrinsic to the phenomenal qualities themselves. It's hard to see how something could have "the same" phenomenal qualities without the qualities having those structures and relationships. But if the system doesn't have the neural features that explain the structure and relationships of the phenomenal qualities, the presence of those structures and relations would be a perfect mystery. To put the same point a different way, our neuropsychology gives us a good empirical hypothesis about the structures that make seeing a red thing the sort of experience it is for humans. Finding those structures in other systems should lead us to attribute similar phenomenal experiences to them. But if we find that other systems lack any relevantly similar structures, scepticism about the similarity, or existence, of their phenomenal experiences seems not unreasonable.

My position is still in need of clarification. In the following sections I shall clarify it by defending it from two kinds of objection. The enemies of qualia, may object that these phenomenal properties are epiphenomenal. The friends of qualia may
object that sub-neural differences could affect the nature of phenomenal experience. I shall develop my position so as to simultaneously counter both allegations.

III.V Epiphenomenalism

Those to whom any talk of phenomenal experience is anathema may argue that the phenomenal properties which I attribute to our neuropsychological states are causally inefficacious. Neuroscientific investigation tells us that the taste sensation of saccharin corresponds to a position in a neural hyperspace that is very sweet, but also very bitter, that is very close on some axes to the tastes of certain other artificial sweeteners, and that is mapped onto other neural spaces in a way that explains inter-modal judgements, such as the notion that saccharin has a rougher taste than sugar. I claim that the event just described has the phenomenal property that characterises the taste of saccharin. The objector argues that this property is causally inefficacious. The causal properties of the neural state would be the same whether or not it had this property. A good way to cash this out is to claim that all the laws linking this state to its effects link it in virtue of properties other than its phenomenal property.

Pace Frank Jackson (1982), I take it to be disastrous for our account if phenomenal properties turn out to be causally inefficacious epiphenomena. I shall spent some time motivating this rejection of epiphenomenalism, as it is critical in determining my final position. Jackson defends the possibility of epiphenomenalism against three objections. The third and most significant objection he considers is the problem posed by epiphenomenalism for our knowledge of other minds. He assumes that we infer the existence of mental states in others on the basis of their behaviour. But if qualia are causally inefficacious, they cannot be implicated in

---

1Jackson conducts his discussion in terms of qualia. A quale is an occurrence of a phenomenal quality, so I take it that nothing in the debate turns on whether I talk of phenomenal properties of sensations, or of qualia associated with sensations.
causing that behaviour. Therefore, runs the objection, there can be no way of inferring the existence of other peoples epiphenomena from their behaviour. Jackson replies that we know that our own brain states generate epiphenomena, and that those brain states also cause behaviour. Knowing this, it is legitimate to argue from the behaviour of others to the causes of the behaviour in their brains, and from these inner causes to the epiphenomena of which they are also the cause. I wholeheartedly agree with Jackson that we can use this pattern of inference to infer that others have phenomenal experiences, but using it is not consistent with supposing those phenomenal experiences to be epiphenomena.

Jackson's "common cause" inference is inconsistent with epiphenomenalism because it assumes that Jackson knows that his own brain states generate qualia. On the epiphenomenalist story, Jackson's belief that he has qualia is not a consequence of his having qualia. He would, ex hypothesi, have all the same beliefs about qualia if he had no qualia. So Jackson can argue from the behaviour of others to the causes of those behaviours in their brains, and out again to the beliefs about qualia that he knows those brain states cause in him. He can legitimately infer that other people believe they have qualia, but he cannot legitimately infer that they actually have qualia. The general problem Jackson faces is in showing how our own belief that we have qualia could be well founded, without making the qualia cause the belief, and thus cease to be epiphenomenal.

Jackson is well aware of the paradoxes of epiphenomenalism, and has urged some of them against Keith Campbell (Jackson 1972). He notices that we could not remember epiphenomena, since our epiphenomena could not cause memories. Nowhere, however, does he explain how we can synchronically believe that we have epiphenomena, although he does say (1982 p133) that we need such an account, and that it must involve the epiphenomena influencing other mental states without influencing anything physical. I find this remark worrying, as it suggests that the price of epiphenomenalism about qualia is wholesale dualism.
One other possibility suggests itself\(^1\). Perhaps the epiphenomenalist thinks 
"awareness" of an epiphenomenon is a separate event from the formation of beliefs 
about that epiphenomenon, and that this awareness has no causal output. On this 
new account we are aware of our epiphenomena, and if there were no 
epiphenomena we would not have this awareness, although all our beliefs, 
including any beliefs we might have about this awareness, would remain the same. 
Let us call this the epiphenomenal awareness position. It allows a nice 
reconstruction of Jackson's common cause argument which seems at first to 
sidestep my objection to the original version. Brain states cause behaviour, beliefs 
and awareness. The awareness causes nothing. But since we have the awareness, 
we can argue from the fact that people display similar behaviour to the fact that they 
have similar brain states, and then, by analogy with our own case, to the fact that 
they have a similar awareness of qualia.

But I cannot see how to use an "awareness" as a premiss in an argument without 
invoking beliefs about that awareness. So if the new argument is to work, Jackson 
needs to show that our beliefs about awareness are reliably related to our 
awareness. The obvious way to do this is with another "common cause" argument. 
The same brain states cause awareness and beliefs about awareness. Hence, beliefs 
about awareness are reliable, although the awareness doesn't cause the beliefs. But 
this will not do. Once again, it calls upon us to argue from the fact that we are 
"aware". I can't imagine what this could mean, unless it means that we use our 
belief in that awareness as a premiss in an argument. But this reading renders the 
argument perfectly trivial. The same brain state causes me to believe that I am 
aware, and to believe that I am aware. Therefore, my belief that I am aware is a 
reliable indicator that I believe that I am aware. We thus return to the original 
problem Jackson faced. The epiphenomenalist has to render his beliefs about 
epiphenomena reliable without breaching the hypothesis of epiphenomenalism.

\(^1\)I am indebted to David Braddon-Mitchell for a spirited defence of this position.
Besides the repetition of the original difficulty, the "epiphenomenal awareness" position faces two other objections. First, it isn't at all clear that we can find anything in our phenomenology that looks like this peculiar "awareness". Imagine trying to determine whether you are genuinely "aware" of your epiphenomena, or merely under the illusion that you are. I suspect that many who suppose that they understand epiphenomenal awareness are using as a model the kinds of conflict that can arise between initial perceptual belief tendencies and the main body of my beliefs. Thus, I can be aware of a visual experience of redness, but form by other means the belief that the thing I am seeing is, in fact, green. In these cases, I can make a distinction between what I am "aware" of, and my beliefs about what I am perceiving. But if I were to lose these kinds of awareness, this would have all sorts of effects on my thoughts and behaviour, and so it isn't causally isolated, as epiphenomenal awareness needs to be. What the epiphenomenalist needs, on the present suggestion, is for me to be aware or not aware of qualia without this awareness or lack of awareness affecting my beliefs and belief tendencies. I doubt that there is any such notion of phenomenal awareness.

Secondly, I suspect that epiphenomenal awareness throws the qualitative baby out with the causal bathwater. The phenomenal experiences which the friends of qualia try to defend are the ones they have beliefs about, can remember more or less vividly, and so on. Beliefs about the nature and existence of our own phenomenal experiences are the absolute paradigm of perceptual beliefs. We are directly acquainted with the qualities of experience involved. But this isn't what epiphenomenal awareness offers us. Instead, we are offered experiences which subjects are synchronically "aware" of, but which they cannot directly have beliefs about or recall after they have happened. The mysterious objects of epiphenomenal awareness are not our everyday phenomenal experiences.

The fact that "epiphenomenal awareness" leads to an implausible account of beliefs about phenomenal experience could easily have been predicted. There are two
ways to form a belief about something. One is to arrive at the belief by inference on the basis of some other beliefs. This was never going to be a plausible model for our introspective beliefs about our own phenomenology. The second is to have a broadly perceptual relation to the state of affairs represented in the belief. As perception or interoception involves a causal link, this has been ruled out ex hypothesi. A last ditch defence of epiphenomenal awareness might admit that we have full blooded perceptual-style beliefs about our phenomenology, but claim that these arise in parallel with the epiphenomenal awareness, creating the illusion that we have beliefs which are directly about our epiphenomenal phenomenology. But then the mysterious epiphenomenal awareness will drop out of the picture, as having such full-blooded perceptual or interoceptive beliefs is enough to constitute having phenomenal experiences.

The paradox of epiphenomenalism is that if there were epiphenomenal qualia, we could not have reliable beliefs about them. Their causal isolation would isolate them epistemically. The truth of epiphenomenalism about phenomenal experience is inconsistent with the existence of evidence for it. The moral I draw from this is that, since we clearly do have reliable beliefs about our phenomenal experiences, they aren't epiphenomenal.

III.VI Neuropsychological Qualia

I now have to develop my account so as to avoid the epiphenomenalist paradox. This can be done in such a way as to simultaneously meet the other challenge which I mentioned above. The friends of qualia may argue that while the conservative inference that other normal humans have the same phenomenal experiences as us is backed by the principle of supervenience, the more radical inference that anything

1 As I have already noted, epiphenomenalism about qualia can be saved by including it in a wholesale epiphenomenalism or dualism. I take it, however, that this is not an option many will find attractive. Certainly Jackson does not seem to envisage it as a possibility.
Appendix III: A Theory of Sensations

neuropsychologically identical to us has the same phenomenal experiences is suspect because changes at sub-neural levels may affect the nature of phenomenal experience. This objection is tightly linked to the problem of causal inefficacy. It is plausible that the neural level, taken broadly to include diffuse electrical and neurochemical effects, is the level at which all state transitions of the mind/brain are determined. If differences at lower levels, such as the change of neural electrolyte considered earlier, change the nature of phenomenal properties, without this changing anything at the neural level, then the change in phenomenal properties has no causal consequences. Allowing this possibility allows phenomenal properties to be causally irrelevant epiphenomena.

I now propose to rule out both the causal inefficacy of phenomenal experience and the suggestion that sub-neural levels could affect phenomenal experience. I claim that the property of being a certain phenomenal experience is nothing over and above the other properties of the sensation at the neuropsychological level. We can best see what this means by recalling just what those other properties are. They are properties that explain, detail by detail, the phenomenologically intrinsic nature of that quality. Its internal structure, how strongly bitterness or sweetness feature in its composition, will turn out to be a consequence of other neuropsychological properties of the sensation state. The relationship of the quality to other qualities, its intrinsic similarities and dissimilarities, are also consequences of other properties of the sensation state. Roughly, I am claiming that the phenomenal experience of saccharin is the experience of a state which is related to all other possible tastes, and to many other sensory modalities, in just the way that the taste of saccharin is. Phenomenal properties are constituted by collections of neuropsychological properties, and do not amount to anything over and above those properties.

This claim that certain phenomenal properties are constituted by neuropsychological properties solves the problem of causal inefficacy. If the claim is correct, then it is not possible for qualia to change while nothing else changes. Any change in the
properties which support the internal structure and similarities of sensations will be detectable by the neurosciences, and, in principle, detectable by the subject and capable of behavioural expression.

The claim that phenomenal properties are constituted by neuropsychological properties will also put an end to the idea that changes at sub-neural levels could change phenomenal properties. They could only do this by altering the neuropsychological properties which collectively constitute those phenomenal properties. Identity at the neuropsychological level guarantees identity of qualia, because it constitutes identity of qualia.

III.VII Some Recalcitrant Intuitions

My final position is interestingly parallel to the position of those functionalists, like Shoemaker (1981) who have asserted that mental states do, indeed, have phenomenal qualities, but that these do not amount to anything over and above the non-phenomenal, functional properties of the state.

Shoemaker and others had the structure of the solution right, but chose too high a functional level for phenomenal qualities to depend on. Working at the traditional supra-neural functional level, they were unable to account for data such as functionally disassociated sensations. Melzack (1973), for example, cites the result that patients who have undergone certain types of lobotomy report sensations identical in quality to their previous, normal sensations of pain. Following their operations, however, they cease to be averse to this "pain". It was this sort of example that originally generated the suggestion that phenomenal properties be identified with lower-level, "physical" properties, rather than behavioural level functional ones. Shoemaker's type of account was also unable to capture the
structural and relational properties of qualities of sensation, except as an arbitrary structure of speech dispositions, like that of Metal Baudelaire. This left them unable to explain these structures and relations, and unable to interpret the utterances that express them. Sensation terms that have some external object associated with them can be taken to refer to properties of those objects, but the resulting assignment of truth values to intra-modal similarity judgements will clash wildly with the views of the speakers. With interoceptive sensations, not even this desperate expedient is available. Thus, the intuition that the qualities of phenomenal experiences might change without any of the properties available to the traditional functionalist changing was well founded.

My account differs from this traditional approach in two main ways. First, I have made phenomenal properties depend upon properties at a much lower level - the neuropsychological level. This allows me to account for such things as functionally disassociated sensations. A neuropsychological pain state may preserve most of its structural and relational properties, while losing its connection to motivation. Certain inter-modal connections would probably be broken in any convincing picture of "painless pain", but enough would remain to account for the subjects insistence that the new, hedonically neutral state "feels the same".

Secondly, I have insisted that to have a phenomenal experience really is to have an internal state that intrinsically possesses a phenomenal quality. Metal Baudelaire passes Shoemaker's test for the possession of phenomenal experience, but he does not pass mine, because his sensation states don't have the right phenomenal properties. It is not enough to claim that there are brute similarities between internal states, there must actually be internal states that have those similarities. Phenomenologically intrinsic similarity claims stands in need of explanation, and validation, and the neuropsychological level is where this explanation and validation occurs. At this level, we are able to give a detailed account of what it is for a sensation to have a quality of a particular sort, and to give truth conditions for
assertions concerning the structure and similarities of those qualities. Assertions of structure and similarity must correspond to the psychological reality.

The intuition that the qualities of phenomenal experiences could change while retaining their individual phenomenological structure and phenomenologically intrinsic similarities is harder to sustain than the traditional suggestion that phenomenal experiences might remain the same while changing their behavioural level functional properties. It is hard to get a grip on the suggestion that the taste of saccharin could change, while remaining equally sweet, equally bitter, equally similar in the same respects to all other sweeteners, equally rough, and related in the same ways to all the other modalities of phenomenal experience. The suggestion that the qualities of experience could change holistically while remaining structurally and relationally identical is perhaps easier to comprehend, partly because it is more abstract. But it can very plausibly be argued that our apparent ability to understand this suggestion is due to a failure to come to terms with just how much the supposedly different qualities have in common with the old ones. I suggest that the answer is everything.

There will, I am sure, be "qualia freaks" who will remain convinced that the essential "feel" of phenomenal experiences has been lost by my account. To someone who has these intuitions, it makes no sense to talk of two things having the same quality in virtue of any degree of isomorphism between them. Quality can never be reduced to structure. Such a person will believe that there can be two total sets of experiential qualities, identical in their internal structure, and in their relations to one another, and yet still different in quality. They differ in a totally isolated property, a difference that makes no difference. The price of such a view, as we have shown, is the epiphenomenalist paradox. We could never have well-founded beliefs about such properties.
I have encountered two main attempts to make these intuitions more plausible. The first is to claim that, while I have shown that causally isolated alterations in qualia are impossible, I have not ruled out the causally isolated total absence of qualia. But this objection fail to appreciate the true nature of my claim. I claim that the structure and relations of the sensation states constitute their qualities. The belief that we could have an identically structurally rich domain of introspective knowledge while being sensationally dead is simply a failure of imagination. The imaginative crutches often used in these cases are the famous "blindsight" experiments. The objector supposes that a "blindsight" person who talked and acted like a normal person would constitute a counterexample to my position. But this is to confuse my position with Shoemaker's. It is precisely the rich structure of introspective knowledge — probably more than we could ever express linguistically — that is lacking in blindsight cases, and which my account guarantees.

The second set of putative counterexamples are versions of Frank Jackson's famous "knowledge argument" (Jackson 1982). In Jackson's original version a neuroscientist is trained from birth in a black and white room via black and white television, and so knows the complete neuropsychology of colour without, allegedly, knowing what the qualia are. I reply that if anyone were able, per impossible, to understand in detail the structure and relations of the complete set of sensory spaces, they would indeed know everything about the sensations. There would, trivially, be a perspectival fact they would not know merely by knowing the neuroscience. They would not know what it was like to have those sensations. But there is no reason in principle why they should not use their knowledge to learn to vividly imagine every sensation, while never having been exposed to the usual

---

1 In the "blindsight" cases, patients are able to act effectively on the basis of information which it seems they must be deriving visually from their environment. They deny, however, that they can see the relevant scenes, or that they are conscious of the information they seem to be using. (Weisenkranz et al. 1974)

2 Of the weakest variety. I do not mean to imply that what they would know from the first person perspective would be different facts about the sensations.
environmental stimulants. They would be like Hume's man who imagines an new intermediate shade, but played out on a grand scale. There may of course, be detailed physiological or developmental reasons why not all sensations can be generated without exposure to the right inputs.

Brian Ellis has put a related, putative counterexample to me in conversation. Suppose we discover that a sensory space of n dimensions has two further dimensions, which have passed unnoticed because every previous input has had zero positions in those dimensions. We set up a pair of special inputs, each of which will register in one of the new dimensions, but have a zero reading in all other dimensions. When we expose someone to these inputs they have, ex hypothesi, two new sensations, and they are different. Call them X and Y. The person who has had the two inputs knows, say, that it was X that was caused by A, and Y by B, rather than vice-versa. The rest of us do not know this, although we have a complete neuropsychology of his sensory system. Therefore what neuropsychology tells us about X and Y leaves something out.

Ellis's example brings out the intuition that underlies "knowledge arguments" very starkly. To make the example bite, he has to insist that the new sensations have no links to other sensory spaces. They are barely different — different from each other, but systematically incomparable with every other sensation. To imagine this we have to imagine a sensory modality that is wholly incommensurable with any of our other sensory modalities. I personally can't do this, but perhaps some who share Ellis's intuitions can. Given the theory I've outlined, I have to say that the two sensations would be phenomenologically barely different. We would be able to discriminate them, but they would have no phenomenal qualities. Thus, our introspective knowledge of them would be a "thin" as our theory of them.

Many philosophers are inclined to deny that they understand what's going on when people claim to have the intuition that bare qualitative difference is possible. I think
that is too harsh. But I think I have shown that the price of sticking with such intuitions is very high. It leads to the paradox of epiphenomenalism. I also hope I've given a rich enough account of phenomenal experience for these intuitions to begin to seem a little attenuated. For those who do still have the intuitions, I can see only one way out. Let's admit that these pure qualitative changes are, in some sense, conceivable. That doesn't mean that they are, in any serious sense, possible. The epiphenomenalist paradox should convince us that they aren't.

III.VIII Conclusions

My theory makes sensations substantive states of the brain, or of other, neuropsychologically identical systems. On this view, a behaviourist or criterialist account of sensations would be as mistaken, in the light of the neuroscientific advances I have speculated about, as Norman Malcolm's infamous criterialist account of dreaming (Malcolm 1959) has turned out to be in the light of advances already made. But as well as wishing to show that Armon-Jones and others are unjustified in assuming that introspective qualities of experience can have no place in a scientific account of emotions, I hoped to give a theory that would allow me to reconstruct as much as possible of the views of MacLean, Izard and Tomkins.

My theory will certainly allow people to recognise their emotional state by the quality of sensation associated with it. We have already noted that it will allow functionally disassociated sensations, and so it will allow emotional sensations to be recognised out of their usual context. The theory will allow the identity of an emotional state to depend on the sensation associated with it, as much as on any other element of the emotion syndrome. To what extent we actually want to make emotion identity depend upon sensations, however, will depend on the details of our emotion theory. I do not see that the cognitive and constructionist theories afford much room for the view that sensations are essential, although, as noted in chapter five, this does not mean that sensations may not frequently play a role in the
recognition of cognitive or constructed states. In an affect-program theory, if an affect-program leads to sensations, then that aspect of the program is certainly essential to it being that very program, although no more essential than any other aspect. However, the looser taxonomies of affect-program's which may be of use in evolutionary theory may not make the sensation essential to membership of an emotion category.

There is one claim of Tomkin's which probably is ruled out by our theory of sensations. At some points Tomkins seems to claim that sensations are "essentially", or "inherently" motivating. If this implies that phenomenologically identical sensations couldn't have different motivational effects in different people then this has to be mistaken. I have argued that it is precisely one of the virtues of our account that it allows for functionally disassociated sensations. There is no reason why a sensation should not change its connections to motivation while retaining its place in the structure of sensations to such an extent as to be phenomenologically identical.
Bibliography


Bibliography


