Empirical Lessons for

Philosophical Theories of Mental Content

NICHOLAS SHEA

KING'S COLLEGE, LONDON

Ph.D. Thesis

2003

Abstract

This thesis concerns the content of mental representations. It draws lessons for philosophical theories of content from some empirical findings about brains and behaviour drawn from experimental psychology (cognitive, developmental, comparative), cognitive neuroscience and cognitive science (computational modelling). Chapter 1 motivates a naturalist and realist approach to mental representation.

Chapter 2 sets out and defends a theory of content for static feedforward connectionist networks, and explains how the theory can be extended to other supervised networks. The theory takes forward Churchland's state space semantics by making a new and clearer proposal about the syntax of connectionist networks – one which nicely accounts for representational development. Chapter 3 argues that the same theoretical approach can be extended to unsupervised connectionist networks, and to some of the representational systems found in real brains. The approach can also show why connectionist systems sometimes show typicality effects, explaining them without relying upon prototype structure. That is discussed in chapter 4, which also argues that prototype structure, where it does exist, does not determine content.

The thesis goes on to defend some unorthodox features of the foregoing theory: that a role is assigned to external samples in specifying syntax, that both inputs to and outputs from the system have a role in determining content, and that the content of a representation is partly determined by the circumstances in which it developed. Each, it is argued, may also be a fruitful way of thinking about mental content more generally. Reliance on developmental factors prompts a swampman-type objection. This is rebutted by reference to three possible reasons why content is attributed at all. Two of these motivations support the idea that content is partly determined by historical factors, and the third is consistent with it.

The result: some empirical lessons for philosophical theories of mental content.

Contents

1

Introduction

Thoughts have content. Some can be true or false, others can be satisfied or left unfulfilled, depending upon how things are in the world. The content of a thought specifies that dependence. It shows how a thought relates to something worldly – how it refers. Furthermore, humans understand one another in contentful terms. We predict and explain behaviour by attributing to people thoughts with certain contents. Thus, everyday understanding of human behaviour is largely underpinned by generalisations framed in contentful terms.

How does this all work? That is one of the most ancient, deep and perplexing questions in philosophy. A philosophical theory of content will be a fundamental part of the answer. That theory will say what makes it the case, metaphysically, that a given mental representation has the content it does. Such a theory will show what kinds of things contents are, such that they can fulfil the functions they do. Thus, it will show how it can be that thoughts refer.

Very many theorists have thought hard about these issues over the years, debating and publishing on them at length. Still, the question has not been convincingly resolved. Views proliferate. Philosophers disagree not only about the solution, but also about the fundamental nature of the problem; or even whether there is a problem at all. Part of the task is to get a clear enough understanding of the phenomenon to be able to pin down what is to be explained. Much of the work in the field makes good use of the traditional tools of philosophy: sharpening our everyday understanding, partly by analysing the

concepts we use; thinking through to fundamental issues; scrutinising theories for consistency, and testing the consequences that follow from them; taking account of relatively abstract considerations, and making connections between problems in seemingly different fields – in short: building theories, at a relatively abstract level. Modern science has added impetus to the philosophical debate. It has produced a wealth of discoveries about the behaviour and cognition of humans and other animals, and about their biological realisation. These findings do not dictate a solution to the problem of mental content, by any means. But they do help philosophers, in two ways: by inspiring new lines of enquiry and by furnishing additional constraints. The range of data that a theory of content must account for, and be consistent with, is much wider now than it has ever been. Accordingly, a theorist of mental content cannot ignore the empirical findings of the sciences of brain and behaviour: experimental psychology (cognitive, developmental, comparative), linguistics, traditional neuropsychology, cognitive neuroscience, ethology and computational modelling ('cognitive science').

Empirical discoveries have been made so quickly that philosophical theories have had difficulty keeping up. That is understandable, since the number of empirical researchers working in fields relevant to mental content outstrips the number of philosophers by several orders of magnitude. Philosophical theories have not yet fully taken account of all the relevant data. This thesis contributes towards that task. There is nothing revolutionary about taking this to be a job for philosophers. But the task is far from complete, and remains crucially important. Furthermore, philosophical progress in the last fifty years suggests that reliance on empirical insights is taking us significantly closer to a full understanding of the phenomena of mental representation, and towards a definitive theory of mental content.

In such a huge field, a thesis cannot cover the philosophical literature comprehensively. Nor is there any possibility of surveying all the relevant empirical data. Accordingly, this thesis attempts a more modest task. It takes a few philosophical issues within the field of mental content, and throws light on them by relying upon a few important empirical findings. The process produces some important empirical lessons for philosophical theories of content. The lessons teach ways that existing theories can be improved, and generate new directions for philosophical investigation.

Given the nature of the project, it will be unsurprising that I assume a relatively strong philosophical naturalism. A minimal naturalism allows space for mental phenomena in the causal order, and accepts that they must be consistent with the findings of the

natural sciences.[1] I work under a stronger assumption: that an adequate theory will allow us to understand mental content in terms of the sorts of entities, laws, properties and relations found in the natural sciences. Thus, a theory of content should show how the content of a mental representation is determined, metaphysically, by its non-intentional properties. I take the following theories of content to frame the terms of the debate: ascriptionist semantics and the intentional stance,[2] conceptual role semantics,[3] informational theories[4] (including those relying upon asymmetric dependence),[5] teleosemantics,[6] interpretational semantics,[7] and empiricist or 'picture' theories.[8]

Commonly, these rival theories are tested by applying them to intuitive cases. That is an important task. A good theory should ascribe the right contents in the range of systems to which it is intended to apply. However, it is less common to ask what content ascription is, such that these theories can explain it. Various topics in this thesis help in this – to understand the nature of the question. Those consequences are drawn out explicitly towards the end, in part II of chapter 6.

How to make progress on such a deep and intractable issue? There is no consensus in the field about where the promising lines of enquiry lie. Accordingly, my strategy is to narrow the focus. I employed two tactics in my research. The first was to look for a reasonably convincing theory for a simple system, to provide a framework. The second was to seek new considerations that could be brought to bear – lines of potentially fruitful investigation for generating new theories, or improving existing ones. In fact, the process of formulating a convincing theory of content for the simple model – connectionist systems – itself generated lines of enquiry for the second task. They arise because the theory of content for connectionist systems advocated in chapters 2 and 3 has some unorthodox features. The remainder of the thesis investigates whether those features are

---

1 Hornsby (1997).

2 Dennett (1981b), (1987).

3 Block (1986).

4 Dretske (1981), Usher (2001).

5 Fodor (1990).

6 Millikan (1984), Papineau (1993).

7 Cummins (1989).

8 Prinz (2002).

objectionable and, if not, whether they could apply more generally. The investigation suggests new forms that a theory of content might take, either as applicable to human cognition as a whole, or to mental representation in some more restricted domain.

In chapter 2, a theory of content for connectionist systems is motivated, spelt-out in detail, and defended. I engage with the debate in the philosophical literature as to whether connectionist systems have contentful states.[9] I make progress by being clearer about the syntax of such systems, improving on existing working assumptions about it. That provides the basis for a convincing theory of content for a certain class of connectionist systems. The theory of content allows the philosophical debate about connectionist content to be resolved decisively. That is an important result in its own right. The theory of connectionist content also provides a framework for the thesis, generating suggestions about theories of content in general, that are explored in subsequent chapters. Most strikingly, the theory explains how entirely new representations can develop syntactically, and how they become contentful as a result.

Chapter 3 asks whether the theory in chapter 2 can be extended to any other systems. I start by explaining the limitations of the approach: it cannot account for conceptual representation or quasi-Fregean sense, nor does it suggest that such systems are fully compositional. However, it can be extended to apply to other kinds of connectionist network, beyond those considered in chapter 2. In particular, I explain how the approach extends to some networks that employ unsupervised learning rules. These networks are plausible models of some biological systems found in real brains. Chapter 3 argues that, therefore, the theoretical approach in chapter 2 may furnish theories of content for such simple biological systems.

One of the most well-established results in cognitive psychology is the existence of typicality effects. Connectionist networks give rise to some such effects. However, they lack the feature usually relied upon by experimental psychologists to explain typicality effects, namely a system of concepts arranged into prototype structures. In chapter 4, I review the evidence for typicality effects and conclude that, even if the effects do arise because of prototype structure, that structure is not content determining. I then explain how the syntactic structure of connectionist systems offers an alternative explanation of some kinds of typicality effects – typicality effects without a prototype. The explanation

[9] Churchland (1991), Fodor & Lepore (1992, ch. 6), Churchland (1993), Fodor & Lepore (1993), Churchland (1996), Churchland (1998), Fodor & Lepore (1999).

does not apply universally, but it should make us more careful in inferring from typicality effects to the existence of concepts structured into prototypes. Furthermore, it cannot be an objection to the connectionist explanation that the syntactic structure of a connectionist system is non-content-determining, because that is also true of prototype structure.

Naturalist theories of content have long been influenced by the image of the classical computer. Whilst this has been a powerful and important model, it also has disadvantages. One is that it encourages the assumption that it is unproblematic to individuate the syntactic states of a cognitive system, leading to the view that the main task of a theory of content is to explain how content should be ascribed to pre-existing syntactic items. My case study of content in connectionist systems shows that this need not be so. In connectionist systems, it was for a long time unclear what characterisation would provide an appropriate syntax. Furthermore, my proposed syntax is only vindicated in the light of the useful content attributions that it allows. The case study shows that theories of the syntax and semantics of a connectionist system must be developed in parallel. That may be true more generally since, in complex organisms, and especially in humans, it is far from clear how to divide up the operations of a cognitive system into syntactic elements. Failure properly to characterise cognitive mechanisms in syntactic terms may have hampered attempts to formulate naturalistic theories of content.

The operation of the human brain is at least partly understood at the cellular level. From the other direction, our everyday practices of interpreting and explaining behaviour allow us to attribute content in a way which is neutral about what entities bear those contents. But it is unclear how to make connections between content attribution and implementing mechanisms. Progress would be made if the syntax of mental processes were better understood. My case study in connectionist content illustrates that it is crucially important to individuate mental syntax more clearly. How should vehicles of content be characterised in a realistic system in which, unlike a classical computer, it is unclear how the implementing mechanism should be divided up into syntactic items? There is a real tension between the philosophical assumption made by many theories of content that the vehicles of content can be presupposed by the theory, and the empirical practice of the brain sciences in which there are only faint hints of what those vehicles might be. I take a particular perspective on the general question of syntactic individuation. It has been suggested that, like content, syntax should be externally

individuated.[10] The case study in connectionist content might be thought to support that idea. In chapter 5, I examine the idea of externalist syntax, differentiating between various strengths of the claim. The major constraints on the nature of syntax derive from the reasons for being realist about mental representation in the first place. Realism about representation entails that token representations with the same content within an individual thinker are intrinsically physically similar. I therefore conclude that, while a system's interactions with the environment may have a role to play in finding an appropriate syntax, the vehicles of content must themselves be individuable in internalist terms. Furthermore, to be useful, a syntactic characterisation must map onto causal processes taking place within the implementing mechanism.

My theory of connectionist content, as formulated in chapter 2, makes the circumstances of a system's development partly determinative of the content of its states. That connection may be relaxed. But the cost is that the motivation is weakened for ascribing content to states of the system at all. Therefore, I embrace that feature as a positive characteristic of the connectionist theory. Chapter 6 examines whether it may be true of other kinds of systems – that content is determined partly by ontogenetic factors. In part I of chapter 6, I give examples where the content we would intuitively ascribe does depend upon developmental circumstances. The examples are drawn both from relatively low level capacities studied in humans and other animals, and from higher level cognitive abilities that are more characteristically human. I also highlight some parts of the recent philosophical literature that lend further support the claim that content could be partly determined by developmental factors.

A general objection to that idea derives from a theoretical reluctance to have any kind of historical factor play a content-determining role in a theory of mental representation. That reluctance surfaces in the form of swampman-type objections. The appropriate answer depends upon what a theory of content is for. A theory of content justified only by reference to intuitive examples and empirical data will not take us far enough. The objection to historical factors is more fundamental. To answer it, we need some picture of why a mode of explanation that ascribes contents exists at all. Part II of chapter 6 canvasses some answers to that question. I don't claim that the list is exhaustive, or that any one of the answers should be preferred. Instead, I show in part III how each of those answers supports a role for historical factors in content determination.

---

[10] Bontley (1998).

The upshot of that discussion is that there is no conclusive objection to historical factors playing a role in determining mental content. Indeed, on some views about the purpose of content attribution, it is clear that historical factors should be partly determinative of content.

Finally, chapter 7 is a brief overview of the progress that has been made in the course of the thesis. It also suggests some avenues for further research.

2

Content in Connectionist Systems

(1) INTRODUCTION

Attempts to explain the behaviour of connectionist systems in contentful terms have hitherto proved unsatisfactory. This chapter begins by arguing that such attempts have been hampered by a bad theory of the syntax of connectionist systems. Everyone agrees that representations in a connectionist system are not localist: they are distributed across a layer of the network. That is to say, each token representation comprises activation levels at all of the nodes in a given layer. However, the content of such patterns of activation is widely assumed, explicitly or implicitly, to derive from the contents to be ascribed to the activation of individual nodes. Thus, the underlying syntactic assumption is that the basic vehicles of content are the activations of individual nodes, combining to produce distributed patterns of activation which are complex vehicles with more complex contents. Many attempts to ascribe content to states of connectionist systems have foundered through making that assumption, explicitly or implicitly; or by failing to be clear about what the basic bearers of representational content are. In this chapter I aim firstly to formulate an alternative proposal for the syntax of connectionist systems. To give it a label, the proposal will be that the vehicles are clusters in activation state space. This improved syntax makes it easier to see how connectionist systems are susceptible to contentful explanation, and I go on to defend a theory of content for some such systems.

Section (2) formulates four minimal desiderata that should be met by a syntax for connectionist systems, and explains why the traditional approach has been inadequate. Naturally, these minimal conditions do not require that a syntax display the

compositionality characteristic of classical computationalism. Section (3) shows that clusters provide a generalisable description of the operation of a network that abstracts away from the details of implementation. A case is made that clusters are content-bearing, in virtue of the role they play in explaining how a network can perform correctly on new samples, outside the training set. I then expound a theory of the content of such clusters. Thus, clusters are shown to be syntactic items, and systems with different connection weights and different numbers of nodes may share the same range of syntactic states, with the same contents.

Paul Churchland has advocated a 'state space semantics' for connectionist systems. Following a protracted debate in the literature with Fodor & Lepore,[1] Churchland offered a substantially new idea – that relations amongst contentful points in the state space of a hidden layer are partly determinative of content (Churchland 1998) – relying on recent empirical work by Laakso & Cottrell (2000) on similarities between systems. In reply, Fodor & Lepore deny that the new proposal has any mileage, or that Laakso & Cottrell's results can be relevant to assessing content similarity (Fodor & Lepore 1999). In section (4), I show how my theory answers Fodor & Lepore's objections. I also argue that my theory is one way of working out Churchland's new approach to state space semantics, and explain the way in which Laakso & Cottrell's results are relevant to content.

Section (5) describes some interesting features of the proposed syntax, including its ability to account for representational development. Section (6) outlines some ways in which the theory may be modified to have broader scope so as to apply, for example, to dynamic networks. Section (7) draws out a series of nice consequences of my syntactic approach, including the potential to account for a broader range of empirically observed phenomena. Finally, section (8) makes a comparison of my approach with other important proposals in the philosophical literature.

(2) SYNTAX IN CONNECTIONIST SYSTEMS

2.1 In Search of a Candidate Syntax

Connectionist networks can be trained to perform a variety of tasks. How do they do it? Some say 'brute associationism': networks contain enough redundancy to fit the training

[1] This debate follows on from Churchland (1991), Fodor & Lepore (1992, ch. 6), Churchland (1993), Fodor & Lepore (1993), and Churchland (1996).

data, but their mechanism of operation cannot be further elucidated. If so, network models cannot help explain how any cognitive task is performed. The sceptic allows that networks can perform some cognitive-type tasks. He can even admit that connectionist models provide existence proofs of the sort of tasks that can be learned from scratch. What the sceptic doubts, however, is that connectionists are able to explain how such performance is achieved. So the charge of brute associationism serves to press the question: how does a network achieve correct performance?

One response to the charge relies on the fact that a trained network can often perform correctly in response to samples outside the training set.[2] That would be mysterious if connectionist models work just by over-fitting a set of training data. How, then, should a network be characterised if its mechanism of operation is to be understood? I don't start by ascribing contents, as many theorists do. Instead, my first move is to get clearer about the vehicles of content. What must the vehicles be like if attributing content to them is going to explain something about a network's mechanism of operation? I use that question to arrive at four desiderata. The desiderata are adequacy conditions on a non-semantic means of specifying vehicles of content, and thus on a syntax for connectionist systems – syntax being a way of individuating vehicles of content in non-semantic terms based on the system's mechanism of operation. This is a minimal sense of 'syntax', and does not import the assumption from classical computationalism that syntactic items combine compositionally so as to display systematicity and productivity.

Of course, there is one sense in which connectionist networks' implementation mechanism is already fully understood. A network's behaviour is determined by its architecture, connection weights and activation function. That characterisation completely describes how the network will react to any input. But it is too network-specific to form the basis of an explanation of correct performance. Networks with the same architecture trained on the same sample set to perform the same task usually arrive at quite different distributions of connection weights, depending upon the initial distribution of random weights and the order in which samples are presented. These different weight matrices show no obvious similarities. But perhaps such networks do have something in common in virtue of which they respond in the same way to the same samples. Such a common description will somehow abstract away from the detail of the

[2] As an illustration, just one example is Elman's recurrent network used to discriminate grammatically correct sentences, Elman (1992).

weight matrices for each network. Thus, the first requirement of a candidate syntax is that it abstract away from the detail of particular weight matrices. But it must still concern mechanism, so it must be determined by the distribution of nodes and weights. That is a second desideratum. A third, closely-related criterion is that the putative syntax should have the potential to be shared between different networks which have been trained on the same task. Only if the syntax is generalisable in this way could it play a role in explaining how different networks trained on the same task all manage to achieve correct performance.

A final desideratum arises because it is not just correct performance in relation to the training set which calls for explanation. Against the charge of brute associationism, I relied on the fact that some trained networks perform correctly in response to new samples, outside the training set. A syntax for connectionist systems might help to explain this remarkable ability.[3] That is the fourth and final goal that I will aim at in formulating a syntactic characterisation.

I should emphasise that the way I have characterised syntax makes no assumptions about the suitability of syntactic states to combine compositionally, so as to give rise to a representational system which is either productive or systematic. Some theorists working with classical computation as a model of cognition define syntax in a way which presupposes compositionality. That would be an inappropriate starting point for my enterprise, since it might beg the question against being able to understand the operation of connectionist systems in representational terms. Thus, I begin with a minimal list of criteria that a syntactic characterisation should meet.

To summarise these desiderata, we would have a very good candidate for the appropriate syntax for connectionist systems if we could characterise states of the system in a way which:-

[3] It is logically possible that the projection to new samples occurs by chance, which would preclude explanation; but that is implausible in the light of the range of cases in which networks do manage to generalise to new inputs.

(i) abstracts away from particular patterns of activation and individual weight matrices;

(ii) captures some underlying property of the network which could be the mechanism by which it performs its task;

(iii) generalises to other networks trained on the same problem; and

(iv) forms part of an explanation of the network's ability to project its correct performance to new samples, outside the training set.

But how, even in principle, could a purely syntactic characterisation have anything to say about a network's behaviour in relation to new samples?[4] Here is a rough preview of my response: syntax could contribute to an explanation provided two conditions are met: (1) the network carries out the same syntactic operations on the new inputs as it carried out in response to inputs from the training set; and (2) there is some property of the new samples themselves, relevant to the task against which the network was trained, in virtue of which the same syntactic operations continue to produce correct results in response to the new samples.[5] Notice that any explanation which adverts to properties of the samples themselves is more than purely syntactic – it must proceed by ascribing content to the system's syntactic states; which in turn lends further support for the claim that the states so-identified are indeed the vehicles of content.

2.2 The Microfeatural Assumption

There is a common assumption about the syntax of connectionist systems, often tacit, which has hampered attempts to ascribe contents to their states. I call it the 'microfeatural' assumption, since it often goes with the idea that the contents of states of a network are complexes of 'microfeatures'. It proceeds by treating the syntax of hidden layers as being roughly the same as the syntax of an output layer.

The output nodes of a connectionist network are usually interpreted as each representing one property or state of affairs of interest. For example, in the NETtalk system (Sejnowski & Rosenberg 1987), each output node represents one possible phoneme. Since the outputs are specified in this way as the goal of training, the trained system

---

4 I.e., with novel input encodings.

5 This claim is explained and substantiated in subsection 3.4 below.

produces more or less discrete outputs – the network responds correctly to a sample by producing activation at only one output node, the one which represents the correct response. In order to investigate the network's mechanism of operation the hidden layer or layers must also be examined. It is extremely common to interpret the nodes of the hidden layer in the same way as the output layer, that is as each representing one property or state of affairs. To find out which property is represented investigators look for what is common to the samples which preferentially activate that node.[6] This can be a simple property, but is more often some complex feature of the samples. The standard assumption is that each node represents such complex 'microfeatures', and that a particular pattern of activation across the hidden layer represents a combination of the microfeatures represented by all the activated nodes (combined in proportions reflecting the different activation levels of the contributing nodes).[7]

Thus, the microfeatural approach assumes that the primitive syntactic items are activations of single nodes. Patterns of activity across a layer of the network are viewed syntactically as combinations of these single node activations.

The microfeatural idea has well-recognised drawbacks. It leads to the ascription of highly complex, often disjunctive contents which look nothing like the contents ascribed by folk psychology to human intentional states. That makes critics suspect that connectionist models entail eliminativism about the terms found in everyday folk psychology.[8] Indeed, some proponents of connectionism embrace eliminativism.[9] Furthermore, content ascription appears to be unavoidably holistic: adding or removing a node, or altering the content ascribed to any single node, alters the content ascribed to every pattern of activation. For the same reasons, the contents ascribed are rarely the same between different systems operating on the same task. So this characterisation does not throw any light on possible commonalities between different networks with different weight matrices performing successfully on the same problem. Finally, it is not clear how the proposal could explain how a network manages to project its classifications to new samples. These unattractive consequences are avoided by the approach to syntax advocated below.

---

6 Berkeley (2000) advocates this practice as a matter of principle.

7 Clark (1996) and Clark (1993), p. 47.

8 Ramsey, Stich & Garon (1990).

9 Two well-known proponents of eliminativism have been Paul M. Churchland and Patricia S. Churchland. For examples of each see P.M. Churchland (1981); P. S. Churchland & T. J. Sejnowski (1989).

(3) THE PROPOSAL

3.1 Clusters as an Abstract Description

I will argue that the vehicles of content in a connectionist system are clusters in state space. State space is way of representing every potential pattern of activation of a given layer of a network. For a layer with *n* nodes, the potential activation level of each node is represented by one dimension of an *n*-dimensional space. Thus, any potential pattern of activation across a hidden layer is represented by a point or vector in state space. Consider a network after training: its weight matrix has been gradually adjusted until the network responds correctly to most of the samples in the training set (reaching 'criterion'). Each such training sample produces its own pattern of activation in a hidden layer, represented by a point in state space. Now consider all the points corresponding to samples to which the network responds correctly. They form an array in state space, with some points close together and others further apart. It is found in practice that, after training, such points are not spread uniformly throughout state space, but often fall into relatively discrete clusters. That is, one effect of training a network is to encourage activation vectors in hidden layer state space to fall roughly into clusters.[10] My proposal is that such clusters are the basic syntactic items in a connectionist system: the vehicles of representational content in the hidden layer or layers.

This proposal does not require any absolute notion of proximity between points in state space. Distances between points are determined by the arbitrary metric of the dimensions of the state space of a given system. But picking out clusters only requires a notion of proximity for the state space of a particular system: some points are close together and others further apart with respect to the overall size of the state space of that system.[11] That is, the distribution of points in state space defines its own metric by which clusters can be ascertained. There will be various methods of determining empirically how

---

[10] E.g., Pollack's (1990) recursive auto-associative memory networks and Elman's (1990) simple recurrent networks, both described in Bechtel & Abrahamsen (2002), pp. 171-187. Much *post hoc* analysis of connectionist networks aims to uncover clustering, and closely related phenomena like principal components and attractor processes. For further examples, see the references to cluster analysis in subsections 3.4 and 3.5 below.

[11] More carefully: with respect to the total size of the volume filled by points representing activation produced by samples to which the network responds correctly.

many clusters are formed and where they lie. And the clusters themselves may have vague boundaries. But the important point is that even a relative measure of proximity in state space will allow points within the same layer to be grouped together based on such proximity.

Proximity in state space is relevant to understanding the operation of a network, for the following reason. Points which are close in the state space of a hidden layer produce similar activations in onward connections: in each of the connections between nodes of that layer and the next (the output layer, or a further hidden layer). These activations are then multiplied by the appropriate weights taken from the network's weight matrix in order to transform the activation vector into a pattern of activation across the nodes of the next layer. Variations in the result of this layer-to-layer transformation depend only on variations in the pattern of activation across the hidden layer (since, after training, the connection weights are kept fixed). So points that are close in hidden layer state space represent similar starting points for the transformation into activation across the next layer.

Of course, in the process of transforming activation from one layer to the next, a network may transform points which are distant in one layer into proximal points in the next. However, nearby points are similar inputs into that transformation process. Speaking metaphorically, the network has to do no work to transform proximal points to proximal points, or distal points to distal points. Indeed, the job of the entire network can be seen in these terms: to take points which are distal in the input layer and, by means of a series of layer to layer transformations, to bring them into proximity, in the correct groupings, in the output layer. The network is trained precisely so as to produce clusters in the output layer. So it is no surprise that the training process also gives rise to clusters in the hidden layer or layers. The only difference is that each output cluster corresponds to a single output node (because the goal of training is so specified). By contrast, the clusters in a hidden layer do not typically correspond to any one node, but are distributed. However, in both cases points which are close in the state space of a given layer are treated as similar by that layer. Conversely, samples which produce similar levels of activation in a given layer are close together in state space. That is, proximity in hidden layer state space is equivalent to similar treatment by that layer. More is said in subsection 3.3 below as to why falling in the same cluster is equivalent to similar treatment by a given layer.

The clustering proposal for syntax is motivated by the experimental practice of carrying out cluster analysis in order to explain a network's mechanism of operation (Clark 1993, pp. 52-53). A simple example will illustrate the point.[12] Sejnowski & Rosenberg (1987) used cluster analysis to investigate the operation of their NETtalk system. They sorted training samples into pairs with their closest neighbours based on distance in hidden layer state space, then grouped the pairs in turn with their nearest pairs, and so on, to produce a hierarchical similarity tree. This analysis uncovered a very interesting phenomenon: the hidden layer drew a broad distinction between vowels and consonants on the way to the network's output classification of samples into individual phonemes. It also made some finer distinctions into categories of phonetic type. It is extremely common to carry out such cluster analysis to see what intermediate classifications are made by the hidden layer. Any network which undergoes supervised learning faces the following problem: to cotton onto the properties of the training samples which count for the task against which it is being trained. The training task is to cluster output layer vectors together by relevant features. At the input layer these samples will typically lie in a discontinuous scatter across state space. On the way to transforming the scatter of points in input layer state space into the correct clusters in output layer state space, the hidden layer(s) may make intermediate classifications. That intermediate classification is uncovered by cluster analysis.

However, the modellers who carry out cluster analysis do not go on to claim that clusters are syntactic items. They obviously think that understanding clusters in hidden layer state space helps explain a network's mechanism of operation, but the claim is never clearly made that such clusters form a syntax for the hidden layer. Sometimes modellers formulate scatter plots for individual nodes of a hidden layer, and take the scatters to be something like syntactic items, indicating what an individual node represents (Berkeley 2000). This is a degenerate form of cluster analysis, based on the assumption that clusters align with individual nodes (in some cases this turns out to be roughly true). Furthermore, it is sometimes found that the same set of samples produce very similar levels of activation at two or more hidden nodes (Dawson & Piercey 2001). Rather than holding that each node individually represents some common property of those samples, it makes more sense to view the cluster which has both nodes as components as a single syntactic item.

---

[12] More examples are given in subsection 3.4 and footnotes thereto.

To avoid confusing generality, I will illustrate my proposal with its application in a particular type of system: a static feed-forward classificatory network. This paradigm will form the basic model for the discussion in the remainder of this section, and the next (section 4). Later sections will show how the proposal can be applied to other kinds of connectionist network.

A static classificatory network consists of a layer of input units, interconnected with one or more layers of hidden units, the last of which feeds forward to a layer of output units. A hidden layer may also receive input from a layer of context units. A non-linear activation function[13] determines a given unit's output as a function of the activations received via weighted connections from all the units in a preceding layer. A set of samples is coded into patterns of input across the input layer. The network's task is to classify this set by activating the correct node of the output layer in response to the coded input. Each output node is taken to represent one property of interest, and that is the goal against which training occurs. That is to say, it is a simple consequence of the way the network is interpreted that the output nodes have representational content — what is correct and incorrect at the output layer is given, and is the standard against which training can take place.[14] It doesn't matter how the weight matrix is adjusted to reach 'criterion' (correct classification of most of the samples). I need only assume that some method of supervised learning is applied (delta rule, etc.): by some means a final weight matrix is arrived at under which the network correctly classifies most of the training samples.

Once the final weight matrix has been obtained, hidden layer state space is plotted by re-presenting samples to the network and measuring the hidden layer activation produced by each, without further adjustment of connection weights (so-called 'wire-tapping' the hidden layer). A point is plotted in hidden layer state space for each training sample which is correctly classified (i.e., almost all of the training set, since the network has reached criterion). The distance in hidden layer state space is calculated between

---

[13] Clusters are only an ineliminable feature in networks that have a non-linear activation function. The behaviour of any system with a linear activation function can be preserved by an appropriate transformation which replaces all distributed representations by local ones: Smolensky (1986), p. 411-413.

[14] From a philosophical point of view an account must be given at some point of how outputs acquire representational content. In subsection 3.5 I indicate how this might go. However, the main purpose of the chapter is to account for how the hidden layer can acquire derived content, assuming that the outputs do represent.

every pair of points, and they are divided into clusters by grouping together points which are relatively close to every other member of the cluster, but relatively distant from other points in state space. These clusters are the syntactic items in the hidden layer or layers. Similarly at the output layer, each node of which is a degenerate cluster in output layer state space: each cluster is a syntactic item. There need be no clusters in the input layer state space.

This subsection has given some reasons why clusters in state space might be an appropriate way to characterise the syntax of a connectionist system. However, the main aim has been to show that clusters satisfy the first two desiderata for a syntax: they are a way of describing the mechanism of operation of a network that abstracts away from particular patterns of activation and individual weight matrices.

3.2 Two Versions of the Cluster Proposal

There are two ways of understanding the idea that clusters are the vehicles of content, either of which will support the claims made in this chapter. The modest proposal is that the basic syntactic items are activation vectors corresponding to the centre of each cluster. Then any potential pattern of activation in the hidden layer is treated as a combination of these basis vectors, and derives its content from them. Each potential pattern of activation is a different syntactic item, as in the microfeatural proposal. The difference is that the basic syntactic states are not single nodes, but rather a number of vectors, each consisting of a different pattern of activation across all the nodes of the hidden layer. A more radical proposal completely abstracts away from patterns of activation. The hidden layer is interpreted in terms of clusters, and all patterns of activation falling within a cluster are type-identified for the sake of explaining the network's mechanism of operation. The network is described as transforming inputs into clusters and then into outputs, without mentioning the actual patterns of activity produced. Compare the syntactic description of a classical computer. Even the most detailed description of the syntax does not advert to the voltages in individual wires. Voltages can be ignored because it is taken as read that voltages between 0V and 1.4V, say, are type-identical (treated as 0), as are all voltages greater than 1.4V (treated as 1).[15] Just as a classical computer's syntax abstracts away from actual voltage levels, the more

[15] The threshold between being interpreted as 0 and 1 (on and off) is determined by the applied voltage at which the semiconducting layer of the computer's transistors switches from insulating to conducting.

radical proposal is that connectionist syntax can abstract away entirely from actual activation levels, by means of clusters. The radical proposal is more powerful, but the modest proposal is sufficient for present purposes.

Where positions within a cluster are significant for downstream processing, disregarding those differences for all purposes would lose relevant information. So, the analogy with the way voltages are partitioned into syntactic types in a classical computer is inexact. If the strong proposal is adopted for connectionist systems, then multiple levels of syntax must be admitted, so that a lower level captures differences within a cluster that are important for downstream processing, when those differences are missed at the higher level (ch. subsection 7.4).

Both versions of the proposal rely on the idea that the syntactic items are distributed across the whole hidden layer. In very simple connectionist systems, those with a linear activation function, a suitable transformation of the weight matrix could produce the same input-output performance via a mechanism in which hidden layer nodes were only ever activated singly (Smolensky 1986, pp. 411-413). However, almost all interesting connectionist networks use non-linear activation functions.

3.3 Other Data Structures?

The reliance on clusters provokes a worry. Might there not be other ways in which the network structures activation in the course of processing, but which are more difficult for us as theorists to discern? Why should clusters have preferential status as the only structure of interest? To answer this, first recall that clustering is the relevant property for outputs. The aim of training is precisely to *cluster together* samples into the outputs that correctly represent the property of the sample which is of interest in the classification. Furthermore, that information exists in the distribution of inputs (otherwise the network would not have enough information to perform the task). The task of the network is to produce this output clustering out of input data in which the classification of interest may be spread out in the input layer state space in a highly disjunctive fashion. How can the network make progress towards this goal? By bringing points into proximity in the state space of a hidden layer. Proximity in state space is the only way, built into the architecture of the system, of getting points to be treated similarly in downstream processing. Granted, a network's weight matrix may be such that disjointed distributions of points are drawn together by a stage of processing. (Indeed, that must occur if the network is to achieve its goal.) But whatever the pattern of weights, proximity in state

space will ensure similar treatment by the next stage of processing. That is guaranteed by the architecture, irrespective of the weight matrix. And proximity in state space is the only way of ensuring similar treatment in the next stage of processing. So proximity between activation vectors is the causally important factor in layer-by-layer progress towards the output goal.

Therefore, in giving a more abstract description of the operation of a connectionist network, it is appropriate to gather together nearby points in state space. Clusters abstract away from particular patterns of activation, but nevertheless reflect the causal operation of the mechanism.

3.4 Clusters as Content Bearing

So far, I have argued that clusters provide an abstract means of describing the mechanism of operation of a connectionist system. To complete the case that clusters are syntactic items, I must demonstrate that they are genuinely content-bearing. Obviously, the mere fact that a network instantiates a particular input-output function will be insufficient to substantiate claims about what is going on in its hidden layer, let alone to attribute content to states of the hidden layer.[16] So, I need some extra purchase on the problem of showing that hidden layer clusters are content-bearing. I derive it from considering how networks manage correctly to classify new samples outside the training set.

Empirical work shows that networks do manage correctly to classify samples which fall outside[17] the set on which they were trained.[18] It is observed that activation produced by these new samples often falls into existing clusters in the hidden layer (Lehky and

---

16 As argued by Haybron (2000), for example.

17 For these purposes it is obviously not enough to have a different real-world sample; it must also produce a different pattern of activation across the input layer when the sample is coded for input.

18 Since that is the whole aim of training a connectionist network, most modelling experiments will test whether they achieve successful generalisation, and very many do (Rumelhart 1989, section 2.3; McLeod, Plunkett and Rolls 1998, p. 61). One example which I mention again below is NETtalk, Sejnowski & Rosenberg (1987). Many more examples may be found in Rumelhart & McClelland (1986), McClelland & Rumelhart (1986) and Bechtel & Abrahamsen (2002).

The same phenomenon is found in dynamic networks. For example, Pollack's (2000) cluster analysis of his recursive auto-associative memory networks trained on syntactic phrase structure trees showed that verb phrases formed one cluster and prepositional phrases another: see Bechtel & Abrahamsen (2002), p. 176.

Sejnowski 1987 & 1988, Hinton 1989, Elman 1991, Dawson & Piercey 2001).[19] So clusters may plausibly form part of an explanation of the network's ability to project its classificatory practice. Describing the mechanism in terms of clusters may show why the network behaves as it does with new inputs. From the point of view of individual patterns of activation, the new inputs have nothing in common with what has gone before. But looking at clusters, the new samples produce activation in the same hidden layer clusters as samples in the training set. When the network is described as transforming samples into hidden layer clusters and onwards into output layer clusters, then it is apparent that the new samples *are* being treated in the same way as some of the samples in the training set.

Thus, characterising the operation of the network in terms of clusters allows us to see it as carrying out the same syntactic operations on new samples as it did on samples in the training set, leading to correct classification of those new samples. This is unlikely to be a matter of chance, so we are driven to look for an explanation: something in virtue of which the same syntactic operations continue to produce correct results in response to new samples. That is to say, the empirically-observed phenomenon I am relying on cries out for the following kind of explanation: the new samples have some property in virtue of which they fall into existing hidden layer clusters, and so cause the network to produce correct responses at the output layer.

That kind of explanation cannot just advert to patterns of activation at the input layer, since new samples differ in their input encodings from anything in the training set. So it must[20] advert to properties of the samples themselves which are relevant to the output classification. The explanation is that the network is able to track some property that is common between a new sample and some of the samples in training set; and that it does so by means of hidden layer clusters. That is to say, the explanation relies on the attribution of representational content.

It is not uncommon for connectionist modellers to say something similar. They explain the ability of their networks to generalise in terms of proximity, in hidden layer state space, of the activation patterns produced in response to new samples (Churchland &

---

19 The converse phenomenon is also often remarked upon, and establishes the same point: that failure of the hidden layer to differentiate into clusters into which new samples may fall explains the fact that the network does generalise to new samples: Clark (1993), pp. 132-135.

20 This is an abductive 'must', since the theoretical possibility remains of a network projecting its classificatory behaviour purely by chance; a possibility that is not demonstrably impossible, but only highly unlikely.

Sejnowski 1992, p. 169). They also use various techniques to encourage hidden layer clustering, with the explicit aim of making the network's performance more generalisable (Mozer & Smolensky 1989).[21] My claim is that, if that form of explanation is valid, it requires that states of the network are contentful.

Cases where the performance of a network fails to project to new samples are precisely those where we wish to say the network has failed to pick up on any relevant property. A network may just have arrived at a 'kludge' solution, fitting the training data and nothing else. Or it may have picked up on the wrong properties of the samples and fail to project for that reason. Consider the connectionist network trained to classify photographs into those which showed tanks from those which did not, the aim being to discriminate camouflaged tanks (Clark 1993, p. 41).[22] When tried on a completely new batch of photos, it failed spectacularly. It turned out that the network had become sensitive only to a difference in light and density between the tank and non-tank photos in the original batch, which had worked as a predictor because the tank photos had all been taken in the morning, and the non-tank photos in the afternoon. So we explain the network's failure to project in this case on the basis that it was sensitive to the wrong properties of the samples. Conversely, when a network's classificatory practice does project, we should say that it has picked up on the right properties. And when there are hidden layer clusters, it is not only the outputs, but also the hidden layer that tracks those properties.

Conversely, where the hidden layer fails to differentiate into clusters, that very fact can be used to explain the network's failure to project its correct responses to new samples (Clark 1993, pp. 132-135, Elman 1991).

So the idea that underpins content attribution is that, in training, the hidden layer has managed to cotton on to some property of the samples which is relevant to the classificatory task. That may just be a rough approximation of an output property.[23] However, it may be a useful intermediate property on the way to making the required output classification. I will give four examples to illustrate the types of properties which

---

21 Similar effects are achieved by the techniques in Dawson et al (2000); and Elman (1991).

22 The researchers at the Stanford Research Institute obviously did not think that the practical failure of this model merited publication. However, its theoretical interest would have. The case is also described by Christiansen and Chater (1992).

23 That appears to be what Laakso & Cottrell's (2000) colour classification networks do.

hidden layer clusters can track.[24] First NETtalk,[25] in which the hidden layer made an intermediate classification of the samples into vowels and consonants, which was relevant to the output task of classifying the inputs phonetically. Second, in Hinton (1989) a network was trained to answer questions about family relationships between individuals. Analysis of the internal structure after training revealed that the network had become sensitive to features like age and nationality, which were not given as training primitives, but which were additional features of the domain relevant to the input-output task. A third example is provided by Elman's (1990) simple recurrent networks trained to predict the next word in a linguistic corpus. The inputs were simply binary codings of words. Nevertheless, cluster analysis of the trained network showed that the hidden layer had organised these words into grammatical and semantic categories: nouns vs. verbs, within nouns into animate vs. inanimate nouns, and within animate nouns into words for humans vs. animals. A final example is Pollack's (1990) recursive auto-associative network trained on sets of syntactic phrase structure trees. After training, verb phrases formed one cluster in the hidden layer state space, and prepositional phrases another – a level of generality not given explicitly in the training data.

In short, to explain the network's ability to project we must advert both to the syntactic characterisation of the hidden layer in terms of clusters, and to the properties of the samples represented by those clusters. That is, the explanation involves thinking about clusters in contentful terms.

To test that these intuitions underpin genuinely contentful ascriptions, observe that the hidden layer clusters can genuinely *mis*represent. When a novel sample produces activation within a particular cluster, but does not have the property common to training samples in that cluster, then the network misrepresents the new sample as having that property. The result will usually be a misclassification at the output layer. Misrepresentation at the hidden layer is described in greater detail in the next subsection (3.5).

To recap, clusters in hidden layer state space are more than merely a convenient abstract description of a network, but must be attributed content, if they are to play a

---

[24] Some of these examples are drawn from dynamic systems. Subsection 5.3 below shows that my theory extends to cover dynamic networks too – for precisely that reason: because the clustering / attractor phenomenon is also found in them.

[25] Sejnowski & Rosenberg (1987).

role in explaining how a trained network manages to project its classificatory practice to new samples. That is, four conditions must be met:

(i) the network must be able correctly to classify some set of samples which differ (i.e., differ in their input encodings) from the training set;

(ii) the new samples must fall into the same hidden layer clusters as samples in the training set;

(iii) they must do so in virtue of sharing some property, of the samples themselves, with the samples in the training set;

(iv) that property must be relevant to the classificatory task.

If those conditions are met, then the ability of a network to project its classificatory practice to novel samples ceases to be mysterious — it can be explained in terms of the properties represented by hidden layer clusters.

It remains an empirical question the extent to which connectionist networks satisfy those conditions. As we saw above, there are good reasons to think that at least some do. Furthermore, connectionist systems and training regimes can be specially designed to encourage the hidden layer to meet those criteria, often with the explicit aim of arriving at projectable classifications (e.g.: extra-output learning in Dawson et al 2000; 'skeletonization' in Mozer & Smolensky 1989; and training in graded batches in Elman 1991).

I am not arguing that all connectionist networks must behave in the way described by conditions (i) to (iv). My claim is just that, if those conditions are satisfied, then the network's behaviour should be described in representational terms, with various syntactic items in the hidden layer (clusters) being ascribed content.

3.5 The Content to be Ascribed to Clusters

Here is a more precise account of how content should be ascribed to clusters. Recall that clusters are first individuated by considering the array of points in state space corresponding to activity produced by training samples which are correctly classified once training is complete.[26] Consider all the samples which produce activation within a particular cluster. Then see what property or properties are common to those samples. Restrict the search to properties that could be relevant (causally or constitutively) to the

[26] I.e., most of the training samples, since the network will have reached criterion.

task of judging whether the input samples have the properties represented by the system's outputs. Do the same for each cluster. Finally, ascertain which properties are distinctive of the samples producing activation in each cluster, in comparison with the other clusters in the layer. This process arrives at the property P kept track of by a particular cluster. The content of that cluster predicates P of the sample being presented to the system. In short:

3.5.1 Content of a Cluster

> Activation of a cluster represents that the presented sample has the property, causally or constitutively relevant to whether the input samples have the properties represented by the output layer, that is common to and distinctive of samples producing activation within that cluster.

So the content is something like: *the currently-presented object has property P.* However, the representation does not have constituent structure corresponding to the object and the property. I use a phrase with subject-predicate structure to describe a complete propositional content which, for the system, is realised by a single state without such structure. The system is doing something rather simpler: feature-placing, which is just to make a claim about a demonstratively presented object, without being able to do so about objects presented in any other way. That the content of a cluster is at the level of complete propositional contents shows up in the fact that clusters can misrepresent.[27] It would be a mistake to think of clusters, in the type of case under consideration, as representing propositional constituents.

Why is this the right way of arriving at the property which figures in the content? The answer is that ascribing just these properties can show what the syntax of the hidden layer achieves on the way to making the output classification. Such contents also explain how the system can project its classificatory practice to new samples. It is uncontroversial that a system's outputs represent various properties of the samples: those which the modeller takes them to represent, and against which the network was trained. These outputs fall into clusters, one corresponding to each output node. Thus, output clusters represent properties of the samples. My claim is that, when the conditions set out in the

[27] The content of a cluster is roughly its truth condition.

previous subsection are met, hidden layer clusters are an intermediate classification, made by the system, on the way to achieving the output classification. As in the output layer, points falling within the same cluster are treated by the network as similar, that is, treated as having the same property. So a hidden layer cluster must represent some property which is common to samples producing activation within that cluster. And obviously only properties relevant to the output classification could be germane to such an explanation of the system's behaviour: which provides the second restriction.

The final part of the content clause mandates that each cluster be ascribed some distinctive property of the samples. The justification for looking for distinctive properties is as follows. By dividing activation patterns into clusters in the hidden layer, the network is both treating them as similar if they fall in the same cluster, and as different if they fall into different clusters. The ascription of content should reflect that fact. So different clusters in the same layer should be ascribed distinctive contents.

The contents of hidden layer clusters must relate to output layer contents, but need not repeat them. The hidden layer may divide training samples up more broadly, or more finely than the output classification, or it may make some orthogonal classification which is relevant on the way to achieving the output task.[28]

My content proposal does not endow networks with original intentionality. Hidden layer clusters are only contentful in virtue of the contents ascribed to outputs. And the output layer is contentful because of the way that a human experimenter interprets the system. The outputs are interpreted as representing certain properties of the samples, and the network is trained until it correctly represents those properties of most of the samples, according to that interpretation. The content of hidden layer clusters derives from their relevance as an intermediate stage in making this contentful output classification. In fact, even the method of ascribing syntax to the hidden layer assumes the system's response to inputs can be judged as correct or incorrect: recall that the hidden layer clusters were individuated by considering the array of activation points in hidden layer state space produced by training samples that the network *correctly* classified after training. In order to reach criterion, a network must have arrived at a weight matrix under which most of the training samples produce a correct response, but not all.

[28] Cf. the four examples of intermediate properties discovered in the hidden layers of actual networks, which are listed in the previous subsection.

According to my proposal, those residual samples that continue to be incorrectly classified are excluded in individuating hidden layer clusters.

Although hidden layer content is merely derivative, it is an important step towards naturalising intentionality to be able to understand the operation of hidden layers in contentful terms, given contents at the output layer. That is because, for a network which has successfully undergone supervised learning, the content to be ascribed to the outputs is obvious, deriving from the intentions of the modeller. Any theory of content will have to respect those output ascriptions. Furthermore, there are promising ways of providing a theory of content for the outputs: see the next subsection (3.6).

According to my theory, a hidden layer can misrepresent a sample as having a property which it lacks, in just the same way that there can be misrepresentation at the output layer. In the latter case, when a sample produces a response at the wrong output node, the network is clearly misrepresenting.[29] The same is true when contents are ascribed to hidden layer clusters by 3.5.1 above. When a novel sample produces activation within a particular cluster, the network is representing the sample as having the property ascribed to that cluster in accordance with the theory. If the sample does not have that property, then the network misrepresents. The usual result of misrepresentation at the hidden layer will be an incorrect output, but not necessarily. A novel sample correctly judged to have one of the properties represented by hidden layer clusters may nevertheless fail to possess the property of the resulting output response (since hidden layer classifications may be more coarse-grained, or orthogonal to, output classifications). Conversely, a network might, by chance, produce a correct output by means of a fortuitously corrected mistake: by misclassifying at the hidden layer, but doing so in relation to a novel sample which happens to have the property represented at the output layer. Such cases are not explicable in terms of the syntax and content of the system, but they are not deductively excluded.

The property forming part of the content of a cluster, as specified in 3.5.1, must be causally or constitutively related to the output properties. That is a substantial constraint. I take it that this restricts the class of properties under consideration in two steps. It brings a restriction to properties that could be found in natural laws and,

[29] The result of misrepresentation during training is that the weight matrix is adjusted to produce a more accurate output — which underlines the point that sometimes outputs misrepresent the properties of the sample.

amongst those, it only considers those which could be lawfully related to the output properties. Not any old conjunction or disjunction of properties will do. Thus, my theory of content for connectionist systems does not attempt on its own to answer all the difficulties faced by theories that rely upon informational connections, as my theory does in this context.

So, consider a toy example where all the points in cluster C1 correspond to red samples, and all the points in cluster C2 to blue samples. Here is the mischievous suggestion: C1 represents P1 = *red or blue or green or yellow,* and C2 represents P2 = *red or blue or green or purple*. All the samples in C1 do indeed satisfy disjunctive property P1, similarly for C2, and P1 and P2 are different. To rule out these kinds of ascriptions I hold that a disjunctive property like P1 is not even a candidate for being causally or constitutively relevant to any output classification, since it does not enter into any causal laws, or bridge laws concerning the constitution of things. Properly to explain why that should be is an entirely different topic in the metaphysics of properties, causation and explanation. It is not a problem peculiar to theories of content, let alone peculiar to my theory of content in connectionist systems. So, for my purposes it is legitimate to presuppose that this question can be answered, and that there is some way of specifying a privileged class of 'natural' properties. This distinction is driven by many philosophical considerations. I will motivate my use of the distinction by mentioning just one.

Goodman (1955) argued that there is an epistemic problem of being able to tell which predicates are projectable. The problem is posed by using predicates which we think are projectable to form a new predicate 'grue' which does not seem to be. Thus, it trades on the fact that conjunctions and disjunctions of projectable predicates need not be projectable.[30] Similarly, Quine (1970) applies the non-projectability of negations to answer Hempel's (1965) paradox of the ravens. Those predicates which are projectable with one another for the purpose of induction, explanation or confirmation are roughly those which enter together into natural laws. A natural law is more than merely a universal generalisation which is necessarily true (Dretske 1977, D. Armstrong 1983). It states the existence of some real connection between the properties themselves.[31] For the

---

[30] Also argued by Quine (1970).

[31] This is to state, not to argue for, a position. Of course, there is a range of different views. Indeed, because it is so difficult to specify what it takes to be a natural law or natural property, some theorists hold that the only useful notion in the vicinity is that of necessarily true universal generalisations.

same reason that predicates which are co-projectable need not be when conjoined, negated or disjoined – conjunctions and disjunctions of natural properties may not themselves figure in natural laws.

My theory does not presuppose any particular theory of natural properties. However, just one answer that would be sufficient for my purposes is found in Millikan's (2000) theory of 'substances' and 'empirical properties'. According to her theory, 'substances' are real existents of which a variety of properties are co-projectable in virtue of some real ontological ground. Central examples of substances are people, other animals and spatiotemporal objects. As Millikan puts it, if I learn that Xavier knows Greek on one encounter, this will hold good on other encounters; and similarly with his having blue eyes, liking lobster and many other features. At a more general level, species and natural kinds are themselves substances. If I learn that one cat likes fish, I can infer that others will; and similarly for many other features of cat kind. This kind of projectability of the properties of a substance will be grounded in some underlying real connection. There are two main types of connection, eternal and historical. *Eternal* substances have instances with a common essence. Usually this essence is some inner nature which is responsible for observed properties, such as in the case of instances of chemical elements. The essence may also arise from the fact that all instances of the substance are formed of similar material by the same natural forces in similar circumstances, for example, asteroids. By contrast, the ground for *historical* substances lies primarily in the fact that all instances of the substance are copied from one another. In this way, the similarity between all printed renderings of Magna Carta is explained by the fact that they are all descended from copying the same original. Similarly, the members of species form a substance – not because they share some inner essence, but because they are historically related to each other. In addition to similarity grounded in copying, there are often conservation mechanisms which eliminate unfaithful copies; like proof-readers in the case of documents and natural selection in stable environments in the case of species (which tends to eliminate unfaithful copies of the genome, except to the extent that genetic mutations give rise to increased fitness).

If a network is trained to keep track of whether samples presented are instances of one of a number of substances, then 3.5.1 would require that hidden layer clusters only be ascribed contents which are real properties related to the ontological ground of the substance. Thus, Millikan's theory of substances and their related properties provides an

example of what is meant by the claim in 3.5.1 that the properties ascribed to hidden layer clusters be appropriately related to the properties registered at the output layer.

Millikan's approach is just one amongst many that would be sufficient to underpin my theory. I can even admit property nominalism, provided there is some way to draw a distinction between predicates that are suitable to feature in causal / constitutive explanations, and those that are not.

3.6 Content of the Outputs

As explained above, my theory of content for clusters in a hidden layer relies upon the system's outputs being contentful. I argued that such contents clearly derive from the intentions of the modeller building the system. Indeed, in order to train the system at all the outputs must be interpreted as contentful, to provide a standard against which error can be judged to allow the weight matrix to be adjusted until the system reaches criterion. For this reason, there is no need to find a way of attributing content, in non-intentional terms, to the outputs of supervised classificatory networks. However, my theory of content for the hidden layer is compatible with many naturalistic theories of content for the outputs.

Informational theories start with lawful causal covariation between a representation and an aspect of what it represents.[32] Once a network is trained, its outputs will causally covary with the properties which they classify, at least amongst the training set. Of course, informational theories have notorious problems with fixing upon plausibly determinate contents. Any such indeterminacy at the output would infect content attribution at the hidden layer. One obvious improvement would be to restrict the content to properties of samples which have been presented to the network in training. That is to follow the general strategy of relying upon actual causal history to constrain content, in the way suggested by Kripke for proper names (Kripke 1972); brought strikingly into focus by Davidson's (1987) swampman example, where he argued that absence of the right actual causal history would deprive his intrinsic duplicate of any contentful states. A causal-informational theory has a good prospect of getting the content of the outputs right, but seems then to trade tacitly upon the intentions of the modeller again, since she selected the samples against which the system would be trained.

---

[32] Dretske (1981), Usher (2001).

Another version of informational atomism is Fodor's asymmetric dependence theory of content (Fodor 1990). Applied in a network, the central claim would be that, amongst the many properties which would cause a particular output node to be activated, there is one correlation upon which the others depend asymmetrically: the former correlations would not exist were it not for the latter, but not the converse. That, too, may be true of the outputs of a trained, supervised network; but the counterfactual is underpinned by the nature of the training regime: it is only because the network was trained so that output 1, say, covaries with property P that the output happens also to covary with properties Q, R and S. The required counterfactual is made true by the intentions of the modeller. So, Fodor's theory would fail to eliminate reliance on the intentions of the human modeller. In fact, the connectionist case nicely illustrates a general worry with Fodor's theory: his counterfactuals do seem to capture something true of contentful states, but the theory is unexplanatory because nothing is said about what underpins the relevant counterfactuals. In the worst case it is the content itself which underwrites our belief in the truth of the counterfactuals. In its application to a network's outputs, it was the training by a human modeller which did the work, again failing to bottom out in the non-intentional. This suggests that Fodor has succeeded in neatly characterising the phenomenon of intentionality, rather than providing a naturalistic theory of what determines content.

Two classes of theory of content could not be applied to outputs: definitional and prototype theories. Both assume that typical representations have constituents which fix their content. Output nodes have no constituent structure on which this could be based. Furthermore, it would be completely against the spirit of my proposal, which attacks the microfeatural idea even within a layer, to see nodes in earlier layers as constituents of output representations. I show below how hidden layer clusters may display some prototypicality effects without having the structure required for a prototype theory of content (subsection 7.3). I go on to argue in chapter 4 that even where representations are formed into the types of structures assumed by prototype theories, that prototype structure is unlikely to be determinative of the content. So definitional and prototype theories are not suitable for determining the content of a system's outputs.

Since output layer contents clearly derive from the intentions of the modeller training the system, the easiest answer to the problem is to be explicitly teleological about outputs, but to remain neutral about the appropriate theory of content to apply to the intentions of the human modeller. That answer is fine for supervised networks, but is

there any prospect that it could extend to a natural system in the real world? In such a case, there would be no human intentions for the teleologist to fall back upon. In chapter 3, I explore the issue of whether unsupervised systems can develop hidden layer clusters in the course of learning. For present purposes I address only the following question: what theory of content could apply to the outputs of such an unsupervised system? A causal-informational account would seem to have good prospects, provided the restriction to actually-encountered samples can be justified (more is said about this in chapter 6). Teleosemantics offers an attractive alternative.

In a realistic system the purposes derived from the human modeller might be replaced by the un-derived purposes which arise from evolution by natural selection. At first pass, instrumental conditioning seems to be a natural analogue of the kind of learning that goes on in a supervised connectionist network.[33] In that case, the reward or error signal may plausibly have content fixed teleologically in terms of its evolved purposes. Whichever way such an analogue is established, my proposal for understanding the internal workings of the trained network in contentful terms will apply. It only requires that the system has developed into its final state in order to fulfil some goal or purpose.[34] Those purposes will provide output contents from which content ascriptions to hidden layer clusters can be derived.

Furthermore, even if there were no natural analogues of supervised learning, my proposal remains valid as a contentful way of understanding the operation of the supervised connectionist networks considered in this chapter. Theoretically, it is an important advance in understanding connectionist models to be able to move from outputs specified by the modeller to a content-involving description of the internal workings of the system.

3.7 Contra Eliminativism

Connectionism has sometimes been a source for eliminativism about folk psychology. Eliminativism is the claim that the contents of human cognitive states will turn out to be

---

[33] There are, however, fierce debates over whether any of the rules for weight-matrix adjustment in supervised connectionist learning could have a natural analogue so as to be a potential basis for instrumental conditioning.

[34] Thus, the disputes about realistic mechanisms of weight adjustment mentioned in the previous footnote do not threaten to undermine my theory of content.

quite unlike anything ascribed in the course of everyday explanations of human behaviour (or, more radically, humans may have no contentful states at all — the very existence of content being an illusory).[35] So what sorts of contents are ascribed according to my theory (3.5.1)? The outputs will represent whatever the network is trained to track. These can be just the kinds of properties which folk psychology takes it that humans represent, and typically are perfectly ordinary. For example, connectionist networks have been trained to discriminate colours, to classify groups of letters phonetically and to differentiate sonar traces of mines from rocks. The contents of hidden layer clusters derive from these output contents – they must be properties of the sample causally or constitutively relevant to the properties represented by the outputs. So the hidden layer will likely represent relatively familiar properties too.[36] Caution is need here, because this is an empirical matter. Further empirical investigation is needed to be sure of the range of possibilities. However, we can gain some insight by looking at the kinds of properties uncovered by cluster analysis. They have been tractable, comprehensible properties. For example, the hidden layer in NETtalk clustered activation into vowels and consonants. Critics might insist that that is an artefact of the technique, because only those kinds of properties were investigated. So, to be sure, the theory of content (3.5.1) must be applied to hidden layer clusters in a variety of connectionist systems, first to check that the proposal works in practice across a range of systems, and second to see what sorts of contents are ascribed to hidden layer clusters.

Pending that empirical investigation, the indications are that, whenever connectionist networks are trained in input-output terms to perform the kinds of tasks which humans undertake, the contents of hidden layer clusters are not alien to folk psychology. What is clear is that there is no pressure from the theory towards complex microfeatures, as there is when it is tacitly assumed that each *node* of the hidden layer must be taken to represent some property which unifies all the diverse samples that produce any activation at that node. In short, if connectionist systems do model some aspects of human cognition, adopting my theory of content avoids the need to be eliminativist about the contents of propositional attitudes.

---

[35] As mentioned above, see for example P.M. Churchland (1981); or P. S. Churchland & T. J. Sejnowski (1989).

[36] It is assumed that complex microfeatural or highly disjunctive properties will not be causally or constitutively relevant to the fact that the samples have some familiar output property — see subsection 3.5.

3.8 The Proposal Assessed: Is it a Syntax?

The burden of this section has been to argue that clusters do indeed provide a syntax for connectionist systems. Here, I will recap by recalling how they meet my desiderata for a syntactic description of the mechanism of operation of a connectionist system (subsection 2.1). They capture an underlying feature of the mechanism by which the network performs its task, in a way that abstracts away from the distribution of connection weights and patterns of activation (desiderata (i) & (ii)).[37] It is a description that generalises to other networks trained to classify the same inputs (desideratum (iii)). It also forms part of an explanation of the network's ability to project its classificatory practice to new samples (desideratum (iv)). By way of confirmation, content is ascribed to clusters in a way that allows misrepresentation, as explained above. The syntax need not display the compositionality that is built into classical computationalism. Minimally, syntax is a way of individuating vehicles of content non-semantically, in a way which is determined by facts about a system's mechanism of operation. All of which adds up to a strong case for viewing the clusters in layers of a connectionist network as being its syntactic items.

From this point of view, the standard way of thinking about connectionist systems arrived at the wrong contents because of working with the wrong syntactic model. However, in rejecting the microfeatural idea, I do not want to jettison the assumption that the syntax is relevant to content ascription, as some theorists do. A completely different approach would be to try to explain the behaviour of the system behaviouristically, ascribing contents in virtue of its input-output behaviour, irrespective of the underlying mechanism.[38] But that approach would be just as inadequate in dealing with connectionist systems, because the problem is just too under-constrained by the input-output characterisation (Haybron 2000). My proposal takes a middle course, according to which the syntactic states of a connectionist system are found at the intersection between a description of a system's internal mechanism and a description of how the whole system behaves in its environment. The syntax is a way of typing states of a system that both maps onto features of its internal mechanism, and explains how the system behaves as it

---

[37] Clusters abstract away from such details of implementation however the proposal is interpreted, but do so most radically if the system's mechanism is described without adverting to patterns of activation at all, but purely as transforming inputs to hidden layer clusters to outputs: see subsection 3.2.

[38] In the way that Ryle (1949), Dennett (1987) and Davidson (1984) do for humans.

does in its environment. Clusters do indeed play this dual role. Chapter 5 of the thesis considers whether that could be true of mental representation more generally.

(4) THE CHURCHLAND – FODOR & LEPORE DEBATE

4.1 The Laakso & Cottrell Test

My theory of content for connectionist systems is broadly within the programme of Paul Churchland's latest version of state space semantics (1998). He is robustly criticised by Fodor & Lepore in their latest salvo in the debate (Fodor & Lepore 1999).[39] In subsections 4.2 and 4.3 below, I explain how my theory avoids Fodor & Lepore's criticisms. I also show that my theory is one way of working-out Churchland's new idea that content attribution depends upon the relations between different points in the state space of a hidden layer (rather than upon the relation between those points and the nodes of the layer). His inspiration was recent empirical work by Laakso & Cottrell (2000), so I start by summarising their results.

Laakso & Cottrell discovered a hitherto unnoticed similarity between different systems trained to classify the same samples. Systems with weight matrices that appear to be entirely different were found sometimes to show this higher-level similarity. Furthermore, their method is applicable between networks with different numbers of input or hidden nodes.

Laakso & Cottrell used their test to compare different networks trained to classify colour samples. The inputs were generated from samples from the *Munsell Book of Color*. 627 samples were chosen distributed equally between red, yellow, green, blue and purple. Thus, each network had 5 output units. The samples were encoded in four different ways, to be suitable for networks having 96, 12, 5 and 1 input units respectively.[40] Networks also varied in the number of hidden units used (in a single hidden layer), from 1 to 10 units. Each network was then prepared as follows:-

---

[39] The earlier exchanges are: Churchland (1991), Fodor & Lepore (1992, ch. 6), Churchland (1993), Fodor & Lepore (1993), and Churchland (1996).

[40] Under some input encodings, reflectance spectra data on the samples were used.

(i) The network was trained on the 627 samples until it correctly classified most of the samples into the five colours represented by the output nodes (reached criterion).

(ii) All the samples were re-presented, without further training, and the hidden layer activation measured in relation to each ('wire-tapping').

(iii) The activation levels produced were considered as points in hidden layer state space. The distances between each point and every other point in state space were calculated.

(iv) The pairs of samples were arranged into rank order based on those distances.

Laakso & Cottrell's method is based on the geometric arrangement of points in hidden layer state space. Any pair of points is separated by a distance in state space, which is a simple scalar quantity. So it is easy to calculate which samples fall close together in state space and which are distant. Laakso & Cottrell's idea is to measure the similarity between two networks by comparing the geometric arrangement in state space of these arrays of points. Each training sample has a corresponding point in the hidden layer of both networks. Laakso & Cottrell's test measures whether points that fall close together in the hidden layer state space of one network are also nearby in the other network. More precisely, Laakso & Cottrell generated a rank ordering of pairs of points for each network, from the closest to the most distant. Since the very same samples produce points in both networks, the rank ordering of inter-sample distances for two networks trained on the same sample set are commensurable. Laakso & Cottrell used a statistical test called the Guzman Point Alienation measure ("GPA") to compare the rank orderings. A perfect match gives a GPA score of +1, reflecting the fact that the pair of samples producing the closest activations in network A are also the closest in network B, and so on down to the most distant. A score of -1 reflects a perfect mismatch.

Laakso & Cottrell discovered a striking result: all the trained networks arranged the samples similarly in hidden layer activation space, irrespective of the number of nodes in the hidden layer (except in the case of networks with fewer than 3 units in the hidden layer). That is, each pairwise comparison between networks produced a high GPA score.[41]

[41] Might all three-layer networks trained on the same problem space have similar hidden layer state spaces, and be bound to be judged similar to each other on the GPA test? L&C's own results show that this was not achieved when the hidden layer has only small numbers of units (1 or 2). Nor is it a necessary feature of

Even in networks with different numbers of nodes in their hidden layer, the geometric arrangement of points in hidden layer state space was substantially similar. Thus, Laakso & Cottrell discovered a deep similarity between networks with quite different architectures and weight matrices, when trained on the same task.

Notice that the method does not rely at all on the absolute position of samples in state space. It looks only at the relative position of pairs of samples. Nor does it require some absolute metric of 'closeness' to be specified — there is no need to compare the 'size' of the state spaces of two networks. Once rank orderings have been obtained, the metrical information is irrelevant. All that matters is that the samples that are closest as measured by network A are also close as measured by the (different) metric of network B. The rank ordering of inter-sample distances in hidden layer state space abstracts away from the individual activation levels, the particular weight matrices, and even from the number of nodes in the hidden layer. The test is applicable between networks with different numbers of hidden nodes, or of input nodes. The only constraint is that the networks under comparison should have been trained on input encodings generated from the same set of real samples. As a result of this generality, it is an extraordinarily powerful test. Hitherto, it was not even clear how to compare networks with the same architecture (numbers of hidden layer nodes) when their weight matrices differed substantially. Laakso & Cottrell's test can do that and more: it applies across networks with quite different architectures. In the next subsection I will consider to what extent Laakso & Cottrell's test can be said to concern content similarity.

4.2 Fodor & Lepore on Churchland's State Space Semantics

Churchland (1998) advocates Laakso & Cottrell's method as a means of judging content similarity between two systems. In doing so, he explicitly gives up his earlier

---

the data: they showed that comparison of input layer state space did not produce high GPA scores. Other models suggest that a particular geometric arrangement of points in hidden layer state space is not an automatic result of networks having been trained on the same task. For example, the results of Dawson et al (2000) show that networks trained on the same task might nevertheless arrive at quite different geometric arrangements of points in hidden layer state space. (Although that paper was not using the GPA measure as a means of comparison.)

microfeatural approach to content.[42] Churchland embraces the new idea that content attribution should depend upon the arrangement of points in activation space:

> 'A point in activation space acquires a specific semantic content not as a function of its position relative to the constituting *axes* of that space [i.e., the old microfeatural idea], but rather as a function of (1) its spatial position relative to all of the *other contentful points* within that space; and (2) its causal relations to stable and objective *macrofeatures of the external environment*.'
> (Churchland 1998, p. 8, his italics)

My theory of content in connectionist systems is an attempt to spell out in more detail how the new approach to content, suggested by Churchland in that passage, might be realised. Since Churchland makes only a very general claim, his position is open to misunderstanding by his critics. My theory makes clear how relations amongst contentful points can play two roles in a theory of content, neither of which leads to regress. First, they allow syntactic items to be individuated: proximal points are realisations of the same syntactic item — cluster — and widely separated points are necessarily realisations of different syntactic items. Second, they constrain the ascription of content: proximal points will fall in the same cluster so must be ascribed the same content, and distant points must be ascribed different contents. However, relations between contentful points do not, in themselves, constitute a contentful level.

Fodor & Lepore (1999) make two broad objections to Churchland's (1998) proposal. First, they take Churchland to be founding content attribution solely in content similarity between systems, and they argue that there can be no notion of content similarity without the possibility of content identity. I consider that argument in the next subsection (4.3). Fodor & Lepore's second line of attack is to argue that, given the resources which Churchland relies on, states of hidden layers with different numbers of nodes could not have the same content. Fodor & Lepore argue that if dimensions of state space can be

---

[42] Nevertheless, Churchland still sometimes writes as if the semantic dimensions of representational space are given by individual nodes:

> 'Assume that a representational model can be characterised in terms of a parameter space, the dimensions of which are those neurons that participate in the model.'
>
> (Churchland & Churchland 2002, p. 907)

thought of as contentful at all (which they doubt), then adding a node to a hidden layer must involve adding a semantic dimension. Since semantic spaces of different dimensionalities must have different contents (their example is: wet and potable ≠ wet and potable and H_2O), they conclude that Laakso & Cottrell's method cannot be a measure of content similarity.

That objection must arise from continuing to think of semantic dimensions as corresponding to hidden layer *nodes*, as they would on the standard microfeatural view. My theory gives up that unhelpful idea. Networks with different numbers of hidden nodes can nevertheless form some or all samples into the same clusters, so that some or all of their hidden layer activation states would have the same contents. The nature of the distance relationships between points in state space does not depend directly on the number of nodes in the layers under comparison. Once it is established that hidden layer clusters are the semantic dimensions, it is clear that Fodor & Lepore's objection misses the mark.

Churchland (1998) was not entirely clear that he had changed his mind about microfeatures: 'I stand by those earlier responses …'.[43] So it is understandable that Fodor & Lepore (1999) should predicate their objections on the standard microfeatural assumption about the syntax of connectionist systems. However, Laakso & Cottrell's work should be seen as instrumental in changing the terms of the debate. Their focus of attention is on the arrangement of points in state space, irrespective of the location or dimensionality of the axes of state space defined by the hidden layer nodes. My clustering proposal is one way of capturing this arrangement of points. Churchland was inspired by Laakso & Cottrell's work to move to the claim that arrangements between contentful points could have a role in determining content. My theory of content for connectionist systems is clearly within that programme. Its merit, in answering Fodor & Lepore's objection, is that it is a more concrete working-out of the way that the arrangement of points in state space is relevant. With a concrete proposal in mind, it is much easier to see why Fodor & Lepore's objection to comparisons being made between hidden layers with different numbers of hidden nodes is misplaced.

Having seen that Laakso & Cottrell's work is an inspiration for Churchland's new programme, within which my theory falls, what of the further claim that Laakso & Cottrell's test is a measure of content similarity? Fodor & Lepore doubt that it can be:

---

[43] Churchland (1998), p. 5.

> ‘ … the sort of criterion Laakso & Cottrell suggest for type individuating brain states across dimensionality differences pretty clearly does not preserve either identity or similarity of the contents of the states.’
> (Fodor & Lepore 1999, p. 400)

They say that Laakso & Cottrell’s method cannot be a test for content similarity, precisely because it *is* applicable between hidden layers with different numbers of nodes. Since, as we have seen, they argue that states of such layers cannot have the same content, any test which allows that they can must be flawed, they conclude. That line of attack is undermined once it is appreciated that networks with different numbers of nodes in a hidden layer may nevertheless have hidden layer state spaces with the same semantic dimensions. Laakso & Cottrell’s method will test for that. But it cannot compare the content of one cluster with another. Instead, it treats of all the clusters in a layer and how they are arranged. This is the feature which is compared between networks.

To see why, observe that between two networks which do show clustering, Laakso & Cottrell’s method measures whether the networks cluster inputs similarly, since it tests whether proximal points in the hidden layer of one network are also proximal in the other.[44] Laakso & Cottrell are careful to restrict the applicability of their test to networks trained in the same environment: networks which respond with the same outputs to the same set of samples.[45] In such cases, given the same syntax, there is no basis for the syntactic states of two networks to have different contents. Thus, a positive result on Laakso & Cottrell’s test does indicate that two networks have the same number of clusters, in the same spatial arrangement in their respective state spaces, and with the same contents.

That makes Laakso & Cottrell’s empirical work very significant. I argued above, in response to Fodor & Lepore’s challenge, that networks with different numbers of hidden

---

[44] Their test does not measure clustering directly, so a high similarity score could be achieved between two networks which happen to arrange inputs into the same geometrical arrangement in the hidden layer without forming them into clusters. However, in networks that do show clustering, the test is a measure of similar clustering.

[45] Laakso & Cottrell (2000), p. 73.

layer nodes can have the same semantic dimensions. Laakso & Cottrell's positive results shows that sometimes they do, in practice.

Caution is needed, though. Their test gives a very strong sufficient condition for states of two networks to have the same content. It is by no means necessary. Two networks may have clusters with the same content, but in different spatial arrangements in state space. In that case, the networks would not be judged similar by Laakso & Cottrell's measure. Furthermore, networks may have some clusters with the same content, and others with different contents. Again, this would result in a low score on the test. What is necessary for same content is that, when the correct theory of content is applied to the clusters (my formulation is 3.5.1), the contents ascribed are the same. Interrelations between points in state space are used to individuate clusters, and the differences between clusters are used in ascribing content, but the nature of the interrelation between clusters does not form part of their content.

As a result, a too-strong reliance on the Laakso & Cottrell test can provoke confusion. For example, Calvo Garzón (2003) objects that Laakso & Cottrell's test will not be positive if points with the same contents are arranged differently in the state space of two networks. On my proposal, different arrangement in state space does not imply different content – content is not fixed as it would be by a conceptual role semantics. Spatial relations between contentful points in state space are crucial to individuating syntax, but do not then form part of the content of these syntactic items. My theory ascribes content at the level of reference, not at anything like the quasi-Fregean level of sense (I suspect Churchland would be uncomfortable with such a strong reliance on the referential level). So I can agree that Laakso & Cottrell's test is not specific to content, but not with the conclusion that the arrangements of points in state space is therefore irrelevant. Such arrangement is crucial to individuating the clusters – the very syntactic vehicles – to which content can then be ascribed.

In short, Laakso & Cottrell's measure is a strong sufficient condition for content similarity, since it also tests for the same arrangement of syntactic items in state space. That is a further interesting aspect of networks' state spaces which can be compared. In the next chapter, I will ask whether this second aspect is also important (see chapter 3, section (4)). For present purposes, it is enough to emphasise that the content attributed to a cluster according to my theory does not mention how that cluster is related to other contentful points.

Thus, the importance of Laakso & Cottrell's work is not in providing a litmus test for content similarity. Rather, it serves two different purposes. Firstly, it highlights the importance of the arrangement of points in state space. I use that as a basis for individuating syntactic items. Secondly, it provides an existence proof: it shows that some networks, trained on the same samples, but with different input encodings and different numbers of nodes in their hidden layer, may nevertheless have the same contentful points in their hidden layers.

So, I would summarise the state of the Churchland – Fodor & Lepore debate about Laakso & Cottrell as follows. In his latest paper, Churchland (1998) is prompted by Laakso & Cottrell to abandon his earlier microfeatural idea. He sets out a new programme for understanding connectionist systems in contentful terms. According to this new framework, the exciting results of Laakso & Cottrell (2000) can be seen to concern content similarity, albeit as providing a too-strong sufficient condition. How that can be so is made clearer in the light of my proposal for characterising the syntax of connectionist systems, and the theory of content to which it is aligned. Understandably, Fodor & Lepore (1999) predicate their objections on the standard microfeatural view about the syntax of connectionist systems. My concrete proposal makes it easier to see why their objections are misplaced.

4.3 Fodor & Lepore's Criticism of Similarity-Based Semantics

The other main line of argument in Fodor & Lepore (1999) is that a test of content similarity is incoherent in the absence of content identity. I agree with Fodor & Lepore that it would make no sense to hold that token states can have similar contents if there were no such thing as content identity. But this is no objection to the connectionist position. Even the most radically holistic conceptual role theorist is not committed to the metaphysical impossibility of content identity. Rather, it is a consequence of his theory that the conditions for content identity are never realised in practice. So he needs some more theoretical apparatus – psychological explanation cannot rely upon content identity alone. To address this additional explanatory issue he needs a theory of content similarity. In other words, he starts with a claim about what it would be for two states to have the same content. He accepts that his theory entails that token states in different systems will not, in practice, have the same content. And he rebuts the implication that this consequence is problematic by formulating an account of content similarity, and showing

how it can do the necessary explanatory work, generalising across states that do not in practice have identical contents.

If Laakso & Cottrell's (2000) test is taken to be necessary and sufficient for content similarity, then relations amongst contentful points are partly determinative of content. As a result, token states in different systems would very rarely, if ever, have the same content. Fodor & Lepore (1999) argue for this conclusion on the basis of the holist nature of the way content would be individuated by that test. Calvo Garzón (2003) adds that systems trained on different sets of samples, or on samples presented in a different order, would almost certainly arrive at different arrangements of contentful points in their respective state spaces. However, none of these arguments shows that content identity is impossible or incoherent within Laakso & Cottrell's framework; rather, they show that it is unlikely to be realised in practice. The limiting case of identity would be achieved if Laakso & Cottrell's measure delivered a correlation of +1. That would be the result, for example, if the test were applied between two networks with the same architecture and starting weights, trained on the same samples the same number of times in the same order. Since those conditions do not arise in practice, content similarity will be a more useful test than identity. But Laakso & Cottrell's test gives no reason to dispose of content identity entirely. Nor should any advocate of Churchland's state space semantics do so.

As explained in the previous subsection, according to my theory Laakso & Cottrell's test is not a necessary condition for content similarity or identity. Relations between contentful points do not play a direct role in individuating content in my theory: they are used to individuate syntactic items, but do not form part of the content to be attributed to those items. To test content identity one must attribute content to the states in question using the criteria in 3.5.1 above. Two states have the same content just in case they are ascribed the same content by the theory. Since the contents ascribed do not advert to conceptual interrelations (or any other analogue of conceptual role semantics), there is no reason in principle why different systems should not have token states with the same contents. Whether or not they do, in practice, will be a matter for empirical investigation. Even if content identity were rare in practice, the theory says what it would be for two different tokens to have the same content. Furthermore, such content identity is clearly not metaphysically impossible. If it is rare, then content similarity can also be employed to do important theoretical work.

Thus, Fodor & Lepore's attack on content similarity is misplaced as an objection to the programme of state space semantics, and clearly does not undermine the roles played by content similarity and identity in my theory of connectionist content.

(5) MODIFICATIONS

5.1 Extension to Other Networks

For the sake of a concrete example, the exposition thus far has been framed in terms of a feedforward network trained to classify some set of inputs. However, my theory of syntax and content for connectionist systems applies more widely. In this section I will show how the theory works in a wider class of cases.

The first assumption to relax is that the network's task is classificatory. The machinery of syntactic individuation applies equally whatever the network's task, even if that task is to produce some behaviour which is not obviously classificatory.[46] The motivation for attributing content also applies more widely. Whatever task a network is trained to perform, it may be able to generalise its correct performance in response to novel inputs. Whenever that occurs, the question will arise how the network has projected its performance outside the training set. As before, an answer is available if the new inputs produce activation in the same hidden layer clusters as found in respect of the training set. And the explanation will involve attributing content to the hidden layer clusters. The type of contents to be attributed will depend upon the task which the network is trained to perform. However, the same overriding considerations apply: the contents must be features that are causally or constitutively relevant to the task which the network performs, and should be common to all samples falling within the same cluster, and distinctive between clusters. In this way, the theory can easily be extended to any static network trained by means of supervised learning to perform any kind of task.

A second challenge is to extend my theory to networks which develop by way of unsupervised learning. The extension to unsupervised networks would require revision to a basic assumption since, according to my theory, the way content is ascribed to a connectionist network derives from the contents ascribable to its outputs. And output contents were understood in terms of the goals against which supervised learning takes place. Substantial work is needed to see if the underlying proposal can be altered so as to

[46] For example, to predict the next word in a word corpus.

be extended to the unsupervised case. In subsection 3.6 above I suggested that, even in the absence of a human interpreter setting goals for supervised learning, the outputs of a biologically-plausible connectionist network could be understood as contentful. The task, then, is to move from the contentful system in which a piece of unsupervised learning is embedded to the ascription of content to new structures which arise from that learning. The prospects are promising. After all, there is very strong evidence that many biological systems deploy distributed representations. In the next chapter I suggest ways in which my theory can be extended to apply to biologically plausible systems (ch. 3, sec. (6)), given the absence of the type of supervised learning assumed in the current chapter.

5.2 Principal Components

Where there are several clusters in a hidden layer, the possibility arises that some of the clusters may be accounted for by simple superposition of others. In that case, the primitive syntactic items are the basic clusters, which are sufficient to account for all the other clusters found in the layer. This raises the possibility of using principal components analysis to uncover the vectors which are sufficient to generate all or most of the points in hidden layer state space. In principle, each principal component corresponds to a basic syntactic item, defining the semantic dimensions which together account for the distribution of points in state space. But care is needed here not to slip back towards the traditional microfeatural approach because, in the general case, when one reparameterisation is available to account for a distribution of points, there will be many other reparameterisations of the same dimensionality. For this reason it is important, in analysing principal components, that most of the components correspond to regions of state space actually occupied by a cluster. It could happen by chance that one of the basic components, found in combination with others in many of the training samples, is not found on its own in any of those samples; in which case that component will be a valid semantic dimension (i.e., syntactic item), despite the fact that it does not align on its own with any clusters in state space. However, that should not be the case in general.

In particular cases there may also be questions about how many clusters are needed to capture the array of points in state space, and of how many components are needed to account for those clusters. However, it does not undermine the theory to find that it leaves open just how best to apply the theory in some particular cases. Nor does the existence of some cases of vagueness undermine the proposal, provided that both the theory, and its application to paradigm cases, are clear. Furthermore, these issues about

the practical individuation of clusters and principal components are likely to become more settled as experience develops of putting the theory into practice. What is abundantly clear on the current state of the evidence is that clusters and principal components can be discovered in very many cases, as illustrated by the success of cluster analysis and principal components analysis of many actual connectionist models.

5.3 Extension to Dynamic Systems

Static feed-forward networks are a basic connectionist architecture, useful for our purposes for their simplicity, making them easier to understand. However, many of the networks which successfully model the most interesting phenomena are dynamic. Because there are feedback connections, the system does not simply react to an input by producing an output response. Rather, it cycles through a series of states, until it settles into some stable condition. To understand such systems, instead of doing cluster analysis, modellers look for attractors in the processes which account for the evolution of the system (Clark 1993, pp. 63-67). One example is Elman (1991a, 1992), where the behaviour of a recurrent network trained to discriminate grammatically correct sentences was explained in terms of the principal components responsible for the network's trajectory through state space.[47] Often, a few principal components are found to account for most of the dynamic behaviour of the system. To extend my syntactic theory to dynamic systems, these attractors or principal component processes must be viewed as syntactic items. Rather than describing the detail of patterns of activation unfolding one into another, the mechanism of the network is described as responding to an input in terms of one or more principal components, which then lead to the output behaviour. That locates the vehicles of content at just the same level of abstraction as clusters in a static network.

Interestingly, the syntax of these dynamic networks is thus described in terms of processes, rather than states. (I use the term syntactic 'item' to cover both states and processes.) Particular states that arise as a dynamic network evolves in its response to some input are not part of the syntactic characterisation. The syntax abstracts away from such transient states, just as it does from individual patterns of activation, relying instead on processes which underlie the evolution of series of states.

---

[47] Illustrated by a trajectory diagram at Elman (1992), p. 162.

This fits nicely with a common worry with thinking about human cognition on the model of a classical computer. The basic syntactic items in a classical computer are indeed states: charges stored more or less transiently in semiconductor gates and other physical media. But it seems much less clear that cognition involves states. From the introspective perspective thought seems to be an unfolding dynamic process. And from the bottom up, neuroscientists study brain processes: patterns of neural firing, etc. Thus, biological systems usually need to be understood dynamically, in terms of processes rather than states (see below: ch. 3, sec. (6) & (7)). My theory of syntax and content for dynamic connectionist systems shows that it is unproblematic to think of a process (i.e., an item extended in time and changing over time) as a potential vehicle of representational content.

Although the microfeatural idea lies behind most theorists' approach to content in static networks, there is no such easy default assumption for dynamic systems. Since the states of the network unfold continually, there is very little temptation to ascribe content to individual patterns of activation in a particular layer. As a result, theorists have to look elsewhere, and often arrive, albeit by a different route, at the same conclusion as I have reached in this subsection: that the vehicles of content in a dynamic network are attractor processes. For example, McLeod, Plunkett and Rolls speculate:

> 'Perhaps the connectionist equivalent of a symbol is a *stable point of attraction in a recurrent network*. Rule-governed behaviour might be the trajectory through a series of attractor basins which a network passes through in performing a task such as processing a sentence.'
> (McLeod, Plunkett & Rolls 1998, p. 276, original italics)

Andy Clark makes the same move, explicitly endorsing the microfeatural idea for static networks, but suggesting that attractor processes may underpin the syntax of dynamic systems.[48] Thus, the theoretical proposal which I have reached via an extension of reasoning about static networks converges with the way dynamic networks are actually understood by connectionist model-builders.

[48] See subsection 8.1 below.

5.4 Processing Topography Analysis

Here is a way potentially to refine the individuation of clusters or principal components. Recall that the absolute value of the metric of inter-activation distance in hidden layer state space is irrelevant, since all we are interested in is the relative distance between different pairs of points in the same layer. Nevertheless, it may be that a given inter-point separation is, in relative terms, more important in some regions in state space than in other regions of the same space. That is, the pattern of weights between one layer and the next may interact with the activation function in such a way that, in some regions of state space relatively large distances make little difference to the response of the subsequent layer, but in other regions relatively small distances between points can make a larger difference to what happens at the next layer. Laakso & Cottrell's measure of inter-network similarity assumes that, within a given layer, relative distances between points are equivalent, no matter in which regions of the layer those points lie. My clustering proposal is not explicit on the issue, but is most easily interpreted under the same assumption. However, that assumption can be relaxed without loss. That is, clusters should be individuated so as to take some account of variations in state space of the importance of inter-point distance to the response of the next layer.

I call this refinement 'processing topography analysis'. The idea is to look at the topography of a hidden layer state space, not just in terms of the basic metric of inter-activation distance, but instead in terms of the differences those distances make to the response of the next layer. That is to view the topography of the hidden layer in terms of the difference it makes to the response of the subsequent layer.

The practicability and utility of processing topography analysis must be tested by empirical application. (Even if it turns out to be an unhelpful suggestion, the basic clustering / principal components idea still stands.) Furthermore, there are many ways in which it could be applied in practice. Here is just one suggestion, to act as an illustration of how the idea might work. Consider all the points in the state space of a hidden layer (not just those that arise in response to some input sample), and for each consider the extent to which small variations away from the point produce large or small differences in response in the next layer. Fix a threshold value for what is to count as a large or small processing difference, and use that threshold value to chart discontinuities in the state space. Vary the threshold value until the discontinuities are such as to divide the state space into a reasonable number of regions. Roughly, a 'reasonable' number of regions will be of the same order of magnitude as the number of clusters found without doing

processing topography analysis. Then reconsider the individuation of clusters in the light of these discontinuous regions, with the prima facie aim of unifying clusters which fall into the same continuous region, and of dividing up clusters which fall across a discontinuity. However, in deciding whether to unify or divide up clusters their putative contents must be considered, so that the final division into clusters can be ascribed relatively simple non-disjunctive contents in accordance with my theory of content, so as to be able to explain the operation of the system.

(6) CHARACTERISATION OF THE SYNTAX

6.1 Syntactic Development

Given the widespread use of connectionist models, a good theory of their syntax and content is of great interest in its own right. For the purpose of my thesis the theory has an even more important role to play. It shows a different way that content might be attributed, and so motivates the discussion of some general issues about theories of content, based on some interesting respects in which the model is non-standard.

Most strikingly, the theory allows for representational development. Theories of content standardly fail to engage with the mechanisms of development, and content has no role to play in accounting for the process of development. Standard theories just take for granted the prior development of syntactic items, with the inter- and mind-world relations needed in order to be contentful. This assumption pushes Fodor towards his implausibly strong conceptual nativism (Fodor 1998).[49] By contrast, my theory explicitly allows for the development of a system's syntax. According to my clustering proposal a connectionist system has no syntactic items when it is assigned random connection weights before training begins. It is only as a result of training that inputs are clustered together in the

[49] Indeed, one recent theory of the distinction between innate and acquired psychological capacities argues that it is criterial that capacities which develop in a way not explainable by any psychological (roughly: contentful) mechanism are innate: see Samuels (2002). If Samuels is right in his characterisation of nativism, then the assumption that Fodor and others make that the acquisition of representational capacities must be presupposed before formulating a theory of content does indeed compel them to the view that such representational capacities are innate. Samuels' test would decide that my theory shows many representational capacities not to be innate (the representational content of the outputs of a network might still be innate, depending upon the case). That is surely an intuitively satisfying result, which could be taken as supporting both my theory of content and Samuels' theory of innateness.

hidden layer. Thus, only after training can a system be ascribed a syntactic mechanism.[50] That is a virtuous consequence. Compare the microfeatural idea. According to that approach, patterns of activation are syntactic items whether or not any training has taken place. They are compounds of the activations of individual nodes. And even when connection weights are set at random, some complex disjunctive microfeatures can be ascribed to each individual node, on the basis of the samples that would cause it to be activated. I take it as a *reductio* of the microfeatural idea that it ascribes contentful states to an untrained connectionist network which consists simply of some architecture of nodes connected by random weights. We have no reason at all to think that the states of such a system are contentful. Yet the rationale for thinking of the hidden layer nodes as each encoding some complex feature of all the samples by which it is causally activated applies equally to the untrained network.

The syntax of a classical computer is also built-in. The system's designer must decide what the primitive representations are to be. On my proposal a connectionist system does not start with any syntactic items. Indeed, it is not even determined how the implementational states (levels of activation) will be partitioned into types. By contrast, as observed above, the way voltages in a classical computer are treated as 0s and 1s is determined by the design of the computer.[51] A common concern about modelling cognition on classical computation is that the primitive representations must all be present at the outset. The system can only learn by forming new complex representations out of these pre-existing components. Of course, whether that is a flaw in the model depends upon ongoing empirical studies of representational nativism. However, it is surely an advantage of my theory that it can account for the development of entirely new primitive representations. Thus, it holds out some hope of fitting together a theory of content with empirical studies of conceptual development, and with any eventual theory of how new representational capacities arise in humans.

In short, my theory of syntax and content in connectionist systems allows liaisons to be made between content attribution and conceptual development. So it acts as an example to motivate an investigation of whether theories of content generally should allow for connections between end-state contents and the mechanisms of learning or

---

[50] Similarly, Rupert (2001) tentatively endorses the idea that clustering in hidden layer state space may be a means of 'coining terms in the language of thought.'

[51] By the transition voltage of the computer's transistors: see footnote to subsection 3.2 above.

development. That investigation is begun in chapter 6 below. Once the case for such liaisons has been made, it is a further task to spell out the way the developmental and learning mechanisms actually found in the psychology of humans and other animals should constrain the contents attributable to the representations to which they give rise. There will be different answers for different mechanisms, so it is a very substantial task, which is only begun in that chapter.

6.2 Role for External Samples in Specifying the Syntax

A second unorthodox feature of my theory is that external-world items have a role to play, not just in determining content, but also in characterising the syntax of a connectionist system. Recall that we uncover the syntactic items by seeing how a layer of a network responds to training samples. These patterns of activity are plotted to form a post-training state space. Those clusters provide the system's syntax, but are not purely intrinsic features of the system, since they depend upon the sample set used to generate the state space of the hidden layer.

Nevertheless, the clusters themselves are not extended entities, nor need externalist properties be used to pick them out. They are regions of nearby points in state space, and closeness in state space is characterised purely internally.[52] The role for external samples is to show which regions of intrinsic similarity are important for contentful ascription. But once they have played that role, they can be dispensed with from the point of view of syntax: the relevant syntactic description can be given purely in terms of intrinsic properties of the connectionist network.

This way of determining syntax may be seen as anything from excitingly unorthodox, to seriously alarming. After all, syntax is supposed to be about the internal workings of a system. On closer examination, it turns out not to be such a weird idea — something which many theorists of content should be happy to accept. The novelty is in doing the work to analyse the issue carefully. Most theories of content just take syntax for granted. It is assumed to be a problem for brain science, without any philosophical interest for a theory of content. In chapter 5 I will argue that the difficulties posed by the individuation of the vehicles of content in realistic systems, like people, are not merely practical. There are reasons of principle to think that it must proceed in tandem with

[52] Even if processing topography analysis is used (subsection 5.4) — that just adverts to the next layer in the processing chain, not to anything external.

content ascription; which means that theories of content must stop taking the syntax for granted. Again, this shows the connectionist case study acting as an intuition pump to thinking about these issues in a new way.

Chapter 5 of the thesis defends the idea that it is legitimate to individuate syntax in the way I have suggested. It also rejects more radical proposals for externalist syntax. For the purpose of supporting my theory of connectionist content in the current chapter, I rely on the fact that my theory does allow the syntax to be characterised in terms of internal properties, so that it is genuinely a feature of the internal workings of a network, irrespective of the environment in which it is found. That is sufficient to sustain the claim that clusters do still count as a syntax.

6.3 Roles for Inputs and Outputs

A third feature of my theory is that both inputs to and outputs from a system play a role in ascribing contents to its states. In chapter 6, I consider whether that may be a desirable feature of theories of content in general. In my specification of connectionist content (3.5.1 above), both inputs and outputs have ineliminable roles to play. Hidden layer contents derive from the contents represented by output nodes. These are ascribed on the basis of the network's purpose — the task against which it was trained to perform. That is an output-oriented way of ascribing content. Hidden layer contents must be causally or constitutively relevant to that output task. However, within that delimitation, the theory looks to sensitivity to inputs to discover what of those properties the network actually represents. The process of determining what is common to and distinctive of samples which give rise to activation within a given cluster is input-oriented. Thus, the theory has a role for both inputs and outputs. In subsection 8.2 and section (9) of chapter 6, I argue that theories of content should allow that both input and output factors play a role in determining content.

The idea that a connectionist network's inputs and outputs are both relevant to content ascription has a nice parallel with empirical work in cognitive neuroscience. A classic example is provided by the electrophysiological studies of Lettvin et al (1959) which discovered that ganglion cells in the frog retina represent bugs. That conclusion was reached only when evidence had been obtained about both inputs and outputs: that increased firing of these cells was generated by bug-like shapes, and that increased firing in turn generated bug-eating behaviour.

6.4 Causal Efficacy

I have claimed that a syntax of clusters is a way of describing the mechanism of operation of a connectionist system. A sample is coded into inputs, which cause a cluster to be activated in a hidden layer, which causes some output node to be activated. However, this description provokes a worry about causal efficacy and causal exclusion.[53] There appears to be a rival causal story to tell, in terms of the activation in each node, its interaction with each connection weight, its transmission to the next node, and its transformation by that node's activation function. Since it is a particular pattern of activation that causes the pattern of activation in the next layer, how can a cluster itself be causally efficacious? If clusters are indeed syntactic items, are they not epiphenomenal?

This kind of worry arises whenever different levels of explanation can be applied to the same physical system. It arises between the treatment of a system by some special science and its description by basic physics; it also arises between different special sciences when they apply to the same things. An important area where the problem seems particularly pressing concerns content attribution in general: why should content be causally efficacious? I address that question in section (10) of chapter 6 below.

For present purposes, it is enough to observe that clusters fall within this general pattern, postponing more general discussion to chapter 6. Clusters provide a completely different level of description from that in terms of patterns of activation. The latter describes the input as some particular pattern of activation (a single point in state space), which causes further activation patterns in subsequent layers, right up to the output layer. The clustering proposal need not mention patterns of activation at all. Some sample is presented to the network (there is no need to advert to its input encoding), which causes a cluster to be tokened in the hidden layer, and subsequent layers, until an output cluster is tokened, representing the property that the network attributes to the sample. In short, this syntactic proposal is just another example of a case where two genuinely different levels of description vie for causal efficacy. It is a philosophical worry which arises in a whole host of areas. Indeed, it should arise for any adequate syntax of any computational system, since the syntax should generalise over some class of lower level causal / mechanistic entities. Exactly the same metaphysical difficulties arise for the contentful

---

[53] Haybron (2000) also considers the issues surrounding causal efficacy in connectionist systems, but from a different perspective. He worries about how information that is stored in a distributed and superpositional fashion can be causally efficacious.

states of human minds given that, on any naturalistic account, they are realised in lower-level physical processes. It might even be counted as a point in favour of my syntactic proposal that it shares these metaphysical difficulties with mental properties in general.

For the purposes of the later discussion, it is important to characterise clustering properties carefully. Clearly, clustering is not a property of any individual node, or even of any layer. It is a property of the whole system: architecture, activation function and weight matrix. Clustering is not inherent in any smaller part. However, the way inputs will be clustered is fully determined by these smaller parts: by the individual nodes, their activation function, and the connection weights by which they are joined. So clustering is a property of the larger entity (the system) which depends upon its components and their inter-relations. However, we look to the samples on which the network is trained to discover which way of carving up the mechanism of the whole system should count as syntactic. Very many intrinsic properties of the whole system are determined by its components and their interrelations. The ones which are to count as syntactic clusters are the ones which meet a certain functional specification: they are caused by inputs and give rise to outputs in a certain way. Thus, extrinsic factors are used to select amongst various intrinsic characterisations of the whole network (this claim is defended at length in chapter 5).

So, in moving to clusters as a syntax for connectionist systems, two separate moves are being made, both of which are important in addressing the metaphysical puzzle of the causal efficacy of clusters. Firstly, we move up a level of description, from properties of parts of the system, to intrinsic properties of the whole system that supervene on the properties of those parts and their interrelations. Secondly, we choose the intrinsic properties of the system which realise a functionalist specification, that of producing certain outputs in response to the set of inputs on which the network was trained.

6.5 Why Go Representational At All?

The problem of causal efficacy is closely related to a particular puzzle in the theory of content: why go representational at all? After all, a syntactic description will fully characterise how the system will behave in response to any input. What does it add to attribute content to these states? I examine that general question in part II of chapter 6. My theory of connectionist content offers an answer specific to its own domain (see subsection 3.4 above). Recall that clusters alone would not explain why a system manages to perform correctly in relation to new samples, which differed in their input encodings

from the samples on which the network was trained. To explain how a network manages to project its classificatory practice we need to see the hidden layer clusters as picking up on some property of the training samples, which projects to the new samples as a means of performing the overall task.

For the sake of the later discussion, I will describe this motivation in slightly more general terms. Firstly, the system displays some stable pattern of input-output behaviour, call it F. Secondly, the system came to realise F because of its past operation in some embedded context, with the system altering internally so as to instantiate F in that context. Thirdly, it is because of some property of the things on which it acts that the system does alter its internal organisation. That is, properties of things in the system's environment are causally responsible for the fact that the system instantiates input-output behaviour F at all. In chapter 6 I undertake a tentative exploration of whether these characteristics generalise.

(7) FRUITFUL CONSEQUENCES OF THE THEORY

7.1 Content from Solving Realistic Action-Based Tasks

My theory of syntax and content for connectionist systems has some interesting and fruitful consequences. This is the exciting bit: having canvassed the considerations which support the theory, we can now draw out some of its ramifications. Some are nice features of the theory, while others are empirical predictions.

My project so far has been to formulate and defend an empirically-informed philosophical theory of content for connectionist systems. From that perspective, the philosophical enterprise is continuous with experimental science. 'Philosophical' just marks the fact that the theory is formulated at a relatively abstract level, and is driven by some conceptually-motivated concerns which have arisen from debates within philosophy, as well as by empirical considerations from a range of disciplines. As an abstract psychological theory, my proposal is susceptible to confirmation or disconfirmation on the basis of the predictions it makes. Some are theoretically central, while other predictions, were they not confirmed, would only require revisions to the theory rather than its total abandonment.

The first attractive feature of the theory is the link that it permits between action and cognitive contents. Hidden layer clusters can develop as content-bearing within a connectionist system set any kind of task. The outputs need not be taken to represent

properties or states of affairs, but may instead be connected up to actions which the system carries out in the world. The success or failure of the actions can form the basis of adjustment of the network's weight matrix so as to improve its responses to a range of inputs. As hidden layer clusters develop, they will represent properties of the inputs which are relevant to deciding between available actions. Thus, the theory provides a model of how contents can arise in the internal mechanism of a system which is just given an action-oriented task.

My theory has another virtue in that those output actions play a role in content determination. This has a close parallel in empirical research. For example, in the early work on the frog's visual system (Lettvin et al 1959, subsection 6.3 above), retinal ganglion cells were found to respond preferentially to small dark moving objects in the animal's visual field. However, in order to elucidate the function of this sensitivity, it was thought very significant that the output effect was to trigger the frog's tongue dart reflex. Together, the two findings strongly suggested that the retinal ganglion cells served the function of catching flies. That is a real-world example where the nature of output actions were thought to play a role in fixing the content of an internal process.[54] My model of connectionist content illustrates why that might be so as a matter of principle.

7.2 Downstream Use of Emergent Clusters

Outputs may represent actions or the classification of inputs into properties. Either way, hidden layer clusters can keep track of properties which differ from those tracked by the output layer. The hidden layer can make a contentful discrimination which differs from that found at the output layer. Sometimes hidden layer clusters are just an approximation to the output properties. Laakso & Cottrell's (2000) networks seem to fall into that class. In other cases, the hidden layer represents something different from the output layer. For example, recall that the hidden layer of NETtalk represented the vowel / consonant distinction, which was just not marked at all amongst the outputs, each of which represented one individual phoneme, unmarked for phonetic type. It will of course be an

[54] In this case the internal states do not discriminate any differently from output actions: there is just a simple circuit from eye to brain to tongue darting. I am not relying upon this example to establish that the cell firing has any particular determinate content – although standard teleosemantic approaches suggest that they represent *flies* – but merely for the methodological observation that attention to output effects has been influential in the past in pinning down functions and contents.

empirical question to discover the extent to which hidden layer clusters have contents that are orthogonal to the contents represented at the output layer, and if so, how different they can be. But there are signs that they often do differ.[55]

Thus, my theory of syntax and content may show how 'representational redescription' is possible (Karmiloff-Smith 1994, Clark & Thornton 1997). Very roughly, representational redescription is the developmental phenomenon whereby an agent that learns correct performance in some domain is able to transfer that ability to cases with a different structure. Connectionist networks can also show this 'sudden leap' in generalisation capacity (Clark 1993, pp. 166-167). The human developmental evidence is standardly explained as follows: children that can perform successfully in some domain by representing inputs in one way can then come to develop the ability to represent the inputs in quite a different way (Karmiloff-Smith 1994). That 'representational redescription' of the inputs makes the child's performance more generalisable.

My theory shows how the process of learning to perform some output task could lead to the development of representational capacities (at the hidden layer or layers) which are orthogonal to those required at the output level. Thus, it might model how a step toward representational redescription may sometimes be achieved. And even if the standard learning paradigm is limited in the amount of representational redescription that it can generate, there are ways of structuring the task or the network to encourage representational redescription to arise.[56]

Once a network has developed a pattern of hidden layer clusters in the context of some output task, those clusters themselves would be available to be used by some other system. That is, suppose that the nodes of the hidden layer connect to some separate system which had no role in the training task. Once clusters have formed, that intermediate classification could be used by another system for a different purpose. It could even provide reinforcing outputs to another multi-layer learning system, from which an additional stage of representational redescription might then be obtained.

Thus, clusters which are active concurrently in two separate subsystems could be associatively connected (see below, ch. 3, sec. (6)). What is needed are interconnections between the two layers in question which are modified by a Hebbian rule. The fact that

---

[55] See the examples in subsection 3.4.

[56] See subsection 3.4: e.g: extra-output learning in Dawson et al (2000); 'skeletonization' in Mozer & Smolensky (1989); and training in graded batches in Elman (1991).

activation in each subsystem has differentiated into clusters provides a basis on which associative methods can operate. This is what Barsalou envisages when he claims that concepts are formed by the development of associative connections between recurrent patterns of activity which occur at the same time in different perceptual systems (Barsalou 2003). So the very fact that intermediate layers in some subsystem cluster inputs together in significant ways can be made use of for other purposes than those for which they arose in that subsystem.[57]

Furthermore clusters, which are distributed representations, can drive the formation of local representations. Projections from a hidden layer which displays clustering can form the inputs to a competitive network.[58] Learning in the competitive network would tend to produce single node outputs corresponding to each cluster.[59]

One of the uses to which clusters could be put is as inputs to a language module: something new with which a word can be associated. Thus, even if a cluster is not used for any other purpose, it could form the basis of a discrimination which is marked in language.

None of the representations discussed here have a subject-predicate structure. (I argued in subsection 3.5 that a cluster represents, and that its content is found at the level of complete propositions – something like a truth condition – without that syntactic vehicle containing the constituent structure of the subject-predicate phrase we use to describe its content.) However, clusters in different systems could be related in a way that forms a new representation which does have constituent structure. Hurford (2002) argues that the ventral and dorsal streams of visual processing (Ungerleider & Mishkin 1982, Milner & Goodale 1995) underpin a basic form of predicate-argument structure. The idea is that locational information in the dorsal 'where' pathway is 'bound' with categorical information in the ventral 'what' pathway to produce a representation with the structure PREDICATE(X). 'Binding' is whatever neural process associates representations in the two pathways so as to be connected, as being about the same object, in a way that is relevant to downstream processing. Hurford claims that auditory processing similarly consists of two streams that can correspond to argument and predicate respectively. Whatever the status of Hurford's detailed claims, it is clear that his model need not

---

57 Notice that this idea gives up on the encapsulation often thought to be characteristic of a modular mind (Fodor 1983, 1985).

58 For competitive networks, see McLeod, Plunkett & Rolls (1998), ch. 6.; Rolls & Treves (1998), ch. 4.

59 For more details, see ch. 3, sec. 6.2.

assume local representation in either dorsal or ventral stream. So clusters in different systems could, in principle, be 'bound' in a way which gave rise to a representation with subject-predicate structure.

Therefore, there are good reasons to think that human cognitive systems have a mixed architecture, with connectionist-type networks sometimes acting as the input to fully compositional, symbolic processing in some kind of language of thought. Indeed, connectionist modellers sometimes experiment with such classical-connectionist computational hybrids (e.g., Dawson et al 2000[60]).

In short, when learning gives rise to clusters in a hidden layer, those clusters are a powerful resource which could be exploited by other cognitive systems.

7.3 Prototype Effects

A large body of psychological research demonstrates that category judgment displays prototype effects. The speed and accuracy of judgments about category membership, whether explicit or operationalised in the context of some task, depend upon how prototypical the stimulus is for its category, or upon how close it is to some key exemplars of the category. Some connectionist networks seem to show prototype effects (e.g. Rumelhart, Smolensky, McClelland and Hinton 1986). Chapter 4 below canvasses the evidence for prototype effects in some detail, and sets out the main experimental results to be explained. In that chapter I argue that at least some prototype effects can arise in connectionist networks, without them having the constituent structure presupposed by prototype theories of concepts. Here is a brief preview, as it applies to the networks considered in this chapter.

My model shows how prototype effects might arise without there being a stored paradigm exemplar or list of prototypical features. When a new sample is encountered, a connectionist network will only be able to project its classificatory ability to that new sample if the sample produces an activation pattern within one of the hidden layer clusters derived from the training set. For a given property represented by a hidden layer cluster, the more prototypical the new sample is as an example of that property, the more likely it is to fall within that hidden layer cluster, and so to have a chance to be correctly classified. So the prototypicality for the network of novel samples will likely predict the

[60] There, a symbolic system was used to drive a network's hidden layer to cluster in a particular way.

network's accuracy in categorising them. Notice here that prototypicality is relative to the class of samples found within the training set, not the actual extent of the underlying real-world class. But in cases where the training set is a fair distribution of real-world samples, the two classes will roughly align.

Furthermore, my theory shows how speed of categorisation can depend on prototypicality. Consider a dynamic network whose behaviour is explained by some set of principal component processes. If a new input is prototypical for the training set it will feed into these principal component processes, allowing the network to settle quickly into its steady state. An untypical input will fit less well with the principal components, so the network will take longer to settle. This is one possible illustration of why, even in the absence of an explicit representation of a category's prototype, a system might nevertheless be able to respond more quickly to new samples which are prototypical.

7.4 Conceptual Nesting

Where there are many clusters in hidden layer state space, there may also be local clusters of clusters. That provides a way that objects may be represented as falling simultaneously under several hierarchically-organised concepts. Individual clusters may represent basic-level categories (e.g., *dog, cat,* etc.), whilst the content of a cluster of clusters is some higher-level category (e.g., *mammal*). This is discussed further in chapter 4.

Elman's (1990) simple recurrent network for learning word dependencies learnt to organise its hidden layer into these kinds of nested categories: nouns vs. verbs, within nouns into animate vs. inanimate, and within animate nouns into humans vs. animals.

7.5 Lesioning

A standard way to investigate the operation of a network is to delete one of its hidden nodes in order to see what kind of behaviour is thereby produced. From the resulting pattern of correct performance and error, conclusions are drawn about the representational role of the lesioned node. Of course, this procedure is motivated by the discredited microfeatural assumption. It has only been useful because hidden layer clusters sometimes align roughly with one or more of the hidden layer nodes. However, a similar procedure could be used to test my theory. The process would be to make a notional lesion, by removing the operation of one of the clusters or principal components. That requires a transformation of the weight matrix so that any pattern of activity with a

component in the direction of the lesioned cluster has that component subtracted from it. Lesioning individual nodes of a network was supposed to correspond to something neurally plausible. My suggestion has no obvious neural correlate. However, it is a means of making predictions about the contentful role played by such clusters.[61]

The empirical prediction concerns patterns of correct responses. When a cluster is notionally lesioned, samples which did fall within that cluster should no longer give rise to a correct output response (at least, no better than chance). And the correct performance of the network should not generalise to new samples which have the lesioned property. The operation of the network in relation to other samples should remain relatively unimpaired.

Where samples fall into a cluster that consists of more than one component, the behaviour should be even more interesting. Consider samples which have the deleted vector as a component. After notional lesioning, they should be treated in just the same way as similar samples which did not have that component. For example, after notional lesioning to remove the component a in hidden layer state space, a sample which would have produced a response in cluster a + b + c should instead be treated in the same way as the samples falling in cluster b + c prior to lesioning. (For example, something treated as *hairy*-and-*pet*-and-*docile* prior to notional lesioning should, after lesioning, be treated in the same way as samples that were just *hairy*-and-*pet*.)

Another kind of lesioning is also relevant. If a node in a hidden layer is not a significant component of *any* hidden layer cluster, then its deletion should make little or no difference to the network's performance. So, my theory predicts that nodes can be deleted without loss if they fail to participate in any hidden layer clusters. Something like this principle has been used to drive a network towards producing generalisable solutions. Mozer & Smolensky (1989) measured the relevance to performance of hidden layer nodes in a trained network. They deleted the least relevant nodes, and then re-trained on the same sample set. This 'skeletonization' process was found to encourage the network to hit upon a solution that went beyond success with the training set, but would generalise to new samples. My theory can explain why 'skeletonization' increases the ability of a network to project its performance to new samples: because it encourages the formation of hidden layer clusters.

---

[61] Thanks to Jon Barton for this suggestion.

(8) COMPARISON WITH SOME OTHER THEORIES

8.1 Clark

There is a large amount of published literature on connectionist content, much of which deploys the misplaced microfeatural idea. The closest to my theory is Churchland's latest version of state space semantics, which was explained above (section 4). However, three other theorists have closely-related views which I discuss in this section, the first and most important being Andy Clark.

Clark advocates the microfeatural approach to the content of the states of static networks.[62] This remains his position in his latest book, even as he endorses the Laakso & Cottrell test as a measure of content similarity (Clark 2001, pp. 66-76), for example:

> '… the activation of a given *unit* in a given context signals a semantic fact (which is hard to describe).' (my italics)

However, when it comes to understanding dynamic connectionist systems, Clark abandons the microfeatural idea, which is hard to make any sense of in the dynamic context. Instead, he suggests a view much closer to my proposal about the appropriate syntax, along the lines of the suggestion of McLeod, Plunkett and Rolls (1998, p. 276) quoted at the end of subsection 5.3 above. He argues that it is an open and interesting position that dynamic analysis of a network (i.e., finding attractors / principal components that account for its dynamic behaviour) can identify the temporally extended physical processes that are the vehicles of representational content in such systems.

Clark is admirably cautious in the scope of his empirical claims. However, he clearly sees the attraction of understanding dynamic connectionist systems in terms of component processes. It is an indication of the allure of the microfeatural idea that Clark does not consider using the same approach for static networks. It is the complexity of the dynamic case which pushes him away from seeing such systems in terms of microfeatures. I would argue that the basic unworkability of the microfeatural idea is just as good a reason for abandoning it, in both domains. Furthermore, if processes are the vehicles of

---

[62] Clark (1993), p. 47; and Clark (1996): 'the explanatory apparatus of future cognitive science will owe little or nothing to the sentential categories of current commonsense'.

content in dynamic networks, that suggests strongly that their analogue, clusters, are the vehicles of content in static networks.

8.2 Tiffany

Tiffany (1999) comments on the debate between Churchland (1998) and Fodor & Lepore (1999). His views are a development of the ideas of Churchland (1998), and so go in roughly the same direction as my theory in this chapter. He argues that Churchland's state space semantics is best seen just as a means of individuating the vehicles of representation. That is in sympathy with my argument that the semantics is better understood once the syntax is clarified; although I argue that Churchland's framework should be filled out into a theory which attacks the syntax and content of connectionist networks simultaneously. I also agree with Tiffany that Churchland should be interpreted as giving up the assumption that the hidden layer nodes are semantically labelled (what Tiffany calls 'stage one state space semantics').

While I endorse Tiffany's aspirations, my theory differs from his view in two substantial respects. My first disagreement is with his claim that the individuation of vehicles of content in the way Churchland suggests is parasitic on some pre-existing theory of content. Tiffany takes the view that syntax is individuated by comparing the interrelations between different vectors in the state spaces of two networks, and that this requires the vectors to be pre-labelled with contents. I disagree. Tiffany seems to think that the Laakso & Cottrell machinery is itself a means of individuating vehicles. My theory is quite different: syntax can be individuated simply by considering points in the state space of a single network, irrespective of what their contents are. Only once clustering has been ascertained do we make comparisons between networks – that is how the Laakso & Cottrell test is used, not to individuate vehicles. Even then, in using Laakso & Cottrell's GPA to compare two networks, there is still no need to label points in state space as contentful. All that need be known is which points correspond to which token samples. Token samples need not be labelled with any of their properties. The test proceeds just by asking whether, for example, samples 1 and 2, which are close together in the hidden layer state space of network A, are also close together in network B's hidden layer state space. As we have seen, provided the networks are embedded in the same 'environment' (i.e.,

given the same task), this will be strong sufficient condition for content similarity.[63] Tiffany laudably seeks to avoid entangling the individuation of syntax with the ascription of content. Nevertheless, in concluding that vehicles of content can only be individuated in the light of a pre-existing theory of content he falls into that trap.

I also endorse a role for external-world samples in individuating syntax, but not in a way which requires a theory of content. Rather, the role of samples is to give rise to an array of points in hidden layer state space, which can then be divided into clusters. The only pre-existing notion of content presupposed for that purpose is that output nodes can be ascribed content (which is accepted all round) – that is needed in order to tell which samples produce a correct response from the network after training, since it is only these samples on which cluster analysis is carried out. Thus, my theory shows that Churchland's state space semantics can be filled out in a way which is not parasitic on a pre-existing theory of content of the points of state space, as Tiffany claims.

My second major disagreement is with Tiffany's argument that the assignment of content to clusters assumes a conceptual role semantics. This follows from taking the Laakso & Cottrell test as a necessary condition for content similarity. Given the way Churchland deploys Laakso & Cottrell (2000) he may indeed be committed to the idea that content ascriptions mention the relation of a contentful point to other points in state space, so that points with that fall in the same relations to external samples may nevertheless differ in content if they fall into different topographical arrangements in their respective state spaces. However, I have argued that no such fine distinctions are needed. My theory attributes contents to clusters which do not advert to things like conceptual roles (interrelations in state space).

By making clear that Churchland (1998) should be read as abandoning his earlier microfeatural idea, and by disentangling issues about syntax and semantics, Tiffany makes important progress towards understanding connectionist systems in contentful terms. The present chapter can be seen as taking up from where he leaves off, and progressing further towards that goal. In particular, I have advocated a concrete theory of content for clusters. My theory shows how cluster-world relations and inter-cluster relations both have a role to play in determining content, and spells out how that should be achieved.

---

[63] See subsection 4.2 above.

8.3 Rupert

Rupert (2001) is not addressed to the Churchland / Fodor & Lepore debate, but does make reference to it. He is concerned with the question of how we acquire new conceptual primitives in a language of thought. His interest in connectionist systems is thus as one way, amongst others, in which human may develop new representational abilities. However, in the process he explicitly endorses the idea of 'identifying the vehicles of content with vectors *or regions* in a state space'.[64] The latter idea ('regions') is the kernel of the idea developed in detail in my theory expounded herein. In subsection 6.1 above I explained in detail how my theory can account for the development of new representational types in a connectionist system. My account is not of the development of concepts, but of new representations with complete propositional contents. Nor are my clusters primitives which combine in the way presupposed by a language of thought. However, my theory is in agreement with one of Rupert's central claims, namely that state space semantics has the resources to account for the development of new representational types in connectionist systems.

(9) CONCLUSION

The syntax of a static connectionist system should be characterised in terms of clusters, in hidden layer state space, of points corresponding to activation produced by samples to which the network responds correctly. Syntax works analogously in dynamical connectionist systems: the vehicles of content are the processes which account for the system's dynamic behaviour in terms of attractors or principal components. In the light of this theory, Laakso & Cottrell's (2000) test is clearly one way to measure content similarity between different networks, being one possible way of ascertaining whether the state spaces of the hidden layers of two different networks display the same geometrical arrangement. My syntactic proposal underpins a theory of content for connectionist systems: the content of a cluster is the property, causally or constitutively relevant to whether the input samples have the properties represented by the output layer, that is common to and distinctive of samples producing activation within that cluster.

My theory can be interpreted or modified to encompass a wide class of connectionist networks. It also has some nice theoretical consequences, and makes

[64] Rupert (2001, p. 517, footnote 31, 2nd para. of footnote).

fruitful empirical predictions. Most importantly, perhaps, the theory highlights some unstated assumptions that lie behind many other theories of content. By providing an alternative picture, it frees up some of the intuitions that derive from the analogy between cognition and classical computation. Thus, it motivates the discussion in my thesis of three further issues, applicable to theory of content more generally, which in turn generate some general constraints on any adequate theory of content.

3

Extending the Account to Biological Systems

(1) INTRODUCTION

Can the approach in the previous chapter be extended to apply to real biological systems? That is the question explored in the current chapter. I start by setting aside features that the theory cannot deal with. It does not apply to representations with conceptual structure – section (2). The theory can only deliver a limited amount of compositionality – section (3). Nor can it account for the existence of contents at the quasi-Fregean level of sense. The topographic arrangement of clusters in state space is an additional level at which connectionist networks can be compared, in addition to the referential contents discussed so far, but that level is not quasi-Fregean sense, for reasons I discuss in section (4). Finally, the account neither relies upon, nor accounts for, the differentiation of representations into beliefs and desires, or any other distinction between contents with different directions of fit – section (5).

Section (6) goes on to look at real biological systems. The first task is to make the case that real brains employ distributed rather than local representations (subsection 6.1). That is becoming increasingly widely accepted. When the brain does employ distributed representations, activity in some region of the brain can be considered as a point in a state

space with very many dimensions. Where a variety of similar inputs produce similar distributed representations, those inputs will form a cluster in that state space – a 'cluster' in exactly the same sense as in the last chapter. The second task of section (6) is to show that behavioural results should be explained in terms of processing over clusters (subsection 6.2). I consider biologically-plausible connectionist models of classical conditioning and show that these should be understood in terms of processing over clusters. The neural processes giving rise to instrumental conditioning are less well described, and the connectionist models correspondingly less biologically compelling. However, here too explanations are likely to proceed in terms of clusters. Furthermore, content should be ascribed to clusters in roughly the same way as advocated in the previous chapter. Section (7) summarises four key features of the theoretical approach in the last chapter. That gives an indication of further areas where the clustering approach might be tried. I give one example.

Section (8) draws the conclusion that one can reject behaviourism without embracing a language of thought for all cognition. A middle ground exists, applicable to some kinds of cognitive task: performance achieved by manipulation of internal representations, where the representations do not have the conceptual structure, or other constituent structure, characteristic of a language of thought. Connectionist models have shown in detail the kinds of tasks that can be performed by these means. My theory of content for connectionist systems vindicates the claim that there is genuinely processing over representations in such cases.

(2) CONCEPTUAL REPRESENTATION

2.1 The Theory in Chapter 2 Does Not Extend to Conceptual Representations

The theory of content for connectionist systems expounded in the previous chapter does not treat of representations with conceptual structure (see ch. 2, sec. 3.5). In subsection 7.2, I suggested a way in which emergent clusters could provide inputs to a conceptual representational system. However, the clusters themselves do not correspond to concepts. Nor do they have constituent structure, with components corresponding to concepts. Rather, they are unstructured representations with complete contents.

According to the theory, the content of a cluster is described using a phrase of the form P(s) (see ch. 2, at 3.5.1). P is the property ascribed, and varies depending upon which cluster of the hidden layer is activated. The subject part, s, is fixed. The system

can only ascribe properties to the sample being presented at that time. So, although the content mentions the current sample – *that sample has property P* – it cannot mention any other individual. This raises a little-noticed issue, if connectionist representations are to be used as input to conceptual representations. It is often assumed that connectionist representations straightforwardly realise concepts. But my theory starts with connectionist representations with complete contents, as do most others. The content of a concept is something incomplete, and thus suitable for putting together with other concepts to form a complete content. Concepts on their own do not make claims. The states of the connectionist systems considered so far do: both in their outputs and in their hidden layers. Perhaps different kinds of connectionist system realise concepts. Then the states of such a system would not have complete propositional contents, but would only make a claim when suitably combined with another concept. That is fine if the two types of system are distinct: those realising concepts and those with complete contents. The difficulty comes, however, with the suggestion that representations in the types of system we have looked at so far can act as inputs to a conceptual system. I endorse that idea (ch. 2, sec. 7.2). So I need to say something about the dual roles. How can a representation whose content is always P(s), where s is fixed, become the predicative constituent of a conceptual representation P(x), where x can vary over a range of individuals?

The answer will depend upon the empirical facts of a particular case. The feature-placing representation must be combined with another representation in a way that is significant for downstream processing. It must be right, from the point of view of the use made of the combination, to see the combined structure as having its own representational content. And that will require an account of why, from the perspective of the way the complete representation is employed, variations in the feature-placing constituent correspond to variations in the property being ascribed. That can be explained case by case (Hurford 2002 is one example). But this extra step is often overlooked, because concepts are elided with complete representations. There is no difficulty in principle with the idea that a given cluster could form part of two representations. Considered on its own, in the context of its output task, it would be a complete representation; and considered as conjoined with other states which are employed together as a single representation for some different purpose it may be a concept. I want to emphasise that, to individuate a concept, something more is required than just identifying a representational system in which variations of syntactic type correspond to variations in the property mentioned in the representation's content. The theorist must also explain

the binding system: a mode of combination of representations whose significance, for the system, is properly understood as corresponding to the combination of concepts into complete structured representations.

By pointing out that there is a step to be made, from states of connectionist systems with complete propositional contents, to connectionist systems that realise concepts, I hope to have made it clear that the representations considered in the previous chapter are not conceptual. That is an important limitation on the theory. The theory does not extend to representations that have conceptual structure. I take it as incontrovertible that some aspects of human cognition employ concepts. The approach to content advocated in the previous chapter cannot be extended to that kind of cognition. This is perhaps the most important limitation on the potential for extending my theory of connectionist content to humans. However, the representational systems of many other animals are not conceptual. Furthermore, very many of the things that go on in human brains are probably not conceptual, but still representational. Section (6) below explores the prospects for generalising the ideas behind my theory of connectionist content to real biological representational systems of this sort.

In excluding conceptual systems from consideration, I obviously miss a dependent phenomenon: that of forming conceptual connections. Many bits of human cognition seem to involve connecting together pre-existing concepts to represent new pieces of information. A person can learn of dogs that they bark, and store that information by forming a new functional connection between her DOG and BARKING concepts. Similarly, concepts can be assembled into hierarchies: dogs as subordinate to mammals. The connectionist systems we have been considering cannot do such things. Dietrich & Markman (2003) use that as an argument against the kind of approach I have been advocating. They argue that cognitive systems must employ symbols that have the structure of a language of thought.[1] Of course, there has been extensive debate about whether cognition must take place in a language of thought.[2] The arguments usually centre on the supposed systematicity and productivity of thought. Dietrich & Markman

---

1 Dietrich & Markman pose the issue as being about 'discrete' representations, and say several different things about what they mean by 'discrete'. What is clear, however, is that they take such representations to be dissociable symbols which combine in a compositional structure.

2 Fodor (1975), Dennett (1981a), Churchland & Churchland (1983), Peacocke (1983, ch. 8), Fodor (1987, appendix: 'Why There Still Has To Be a Language of Thought'), Sterelny (1990), Smolensky (1991).

instead emphasise the encoding of new pieces of information using existing representational resources. The systems I have been considering do not encode new information using existing representational resources. In order to be able to represent something new, the system must develop a new representational type. The power of the connectionist approach is that it shows how such new types of representations can develop. That development can start from existing representational resources (at the input and output layers). But it does not consist in connecting existing representations together to make a new claim, which is what happens when functional connections are formed between concepts. That does not exclude connectionist systems from playing a role in that process. However, it emphasises the fact that something must be added to the connectionist approach, if it is to be part of an account of human conceptual cognition. Models based on a language of thought can explain how a thinker represents something new, without having to develop a new basic representational type. They do so by combining existing concepts into a new claim. However, the language of thought theorist has the converse problem: he leaves it unexplained how new basic representations develop. The complementary strengths and weaknesses suggest that the two models could usefully be combined in some domains.

2.2 Content is Not Determined by Constituent Structure

As just discussed, some concepts may be functionally interconnected in information structures: DOG - SUBORDINATE TO – MAMMAL. Concepts may also, perhaps, be formed out of others: BACHELOR = UNMARRIED & MAN. Either way, concepts have constituent structure. It is very common to think that a concept's constituent structure has a role to play in determining its content. (The content of a concept is not found at the level of complete propositions – a claim or truth condition. It is the systematic contribution which that concept makes to the complete contents of the representations of which it forms a constituent.) That idea is shared by those who rely on concepts having: classical definitions (Jackendoff 1989), prototypes (Rosch 1978), possession conditions (Peacocke 1992), associated theories (Murphy & Medin 1985), or other kinds of conceptual role (Block 1986). My clusters in connectionist systems do not have that kind of reference-determining structure. They do have some associated structure, since they are found in a

topographical arrangement in state space.[3] However, their structure in state space is used to individuate vehicles, not to determine content. So there is just no prospect of using constituent structure to determine reference. In the next chapter – on typicality effects – I argue that prototype theorists have made a mistake in giving a reference-determining role to prototype structure. Laurence and Margolis (2002b) argue that many other kinds of structure associated with concepts is non-content-determining. Whether or not that is true of concepts in general, it is clear that my approach to the content of complete representations in connectionist systems does not give a reference-determining role to constituent structure.

(3) COMPOSITIONALITY

3.1 The Compositionality of Thought

The most common philosophical complaint against connectionist systems is that they fail to account for the productivity and systematicity of thought. I don't propose to recapitulate the compositionality debate about connectionist systems in general, since it has been explored extensively elsewhere,[4] and because I accept that the kinds of connectionist system considered in the last chapter are not fully compositional. The purpose of this section is to get clear about the phenomena, and to explain why this limitation, although important, still leaves open many aspects of cognition as potentially susceptible to the connectionist approach. Accordingly, the range of cognitive phenomena which require some kind of language of thought may be relatively narrow. That range may stretch little beyond the use of language and the employment of linguistic abilities in the internal mental economy – although those, of course, are some of the most important and wide-ranging aspects of our cognitive life.

---

[3] There may also be clusters of clusters: ch. 2, ss. 7.4; and there may be semantically-relevant relationships between some of the clusters, where basic clusters are principal components of other clusters: ch. 2, ss. 5.2. However, this structure is neither fully compositional (see subsection 3.2 below), nor content determining. Thus, it is nothing like conceptual structure.

[4] Fodor & Pylyshyn (1988), Smolensky (1988), Fodor & McLaughlin (1990), Smolensky (1991), Smolensky (1995).

The basic phenomenon is systematicity: a thinker who can think *Fa* and *Gb* is not conceptually debarred from thinking *Fb*.[5] That is certainly true of what humans can say in public language. And it seems to be true of the kinds of thoughts that are expressible in words. How is this striking systematicity to be explained? The only plausible candidate relies on thought having a compositional structure. Since the thoughts *Fa* and *Gb* are both composed of concepts, those concepts are available for re-use, as is an appropriate mode of combination. That is, having the first two thoughts ensures the thinker has all he needs to think *Fb*: he has the concept F, the concept b, and a mode of combination corresponding to first-level predication.

Opponents of connectionism also cite the productivity of thought. Productivity is our seeming ability to form an unbounded number of different thoughts. One basic way we do this is using the connectives of propositional logic. A thinker who grasps logical conjunction (&) and can entertain thoughts *p* and *q* would seem to be able to think *p & q, p & p, p & q & p,* and so on, until she runs out of representational resources. Again, that is true of the things that humans can say in language. Productivity is also explained by compositionality, but notice that it is a different kind of compositionality from that relied on to account for systematicity in the last paragraph. Productivity requires a means of concatenating complete representations. Systematicity required that representations be structured out of constituent concepts. The compositionality of conceptual structure does not lead to an unbounded number of thoughts – with n singular concepts and m first-level predicate concepts a thinker is only ensured (n x m) complete thoughts. For representations with constituent structure, there may also be compositionality at the level of concepts: the ability to form new concepts out of existing ones. Thus, a thinker who can think Fx and Gx, and is possessed of a mode of combination of predicative concepts, can thereby think (F&G)(x). The ability to form concatenative concepts is obviously closely related to the ability to concatenate complete representations. They may even be the same thing: perhaps (F&G)(x) is equivalent to (Fx)&(Gx). However, the basis for productivity is concatenation, which is different from the basis for systematicity, namely conceptual structure. The latter is clearly absent in the kinds of systems considered in chapter 2 above. As we will see in the next subsection, clustering in state space may give rise to and explain certain kinds of concatenative abilities in connectionist networks. How

[5] F and G are schematic letters for first-level predicative concepts, a and b for singular concepts. *Fa* (in italics) refers to the content arrived at by predicating concept F of concept a.

much of a drawback is it that connectionist representations lack the kind of compositional structure that would account for systematicity?

Smolensky (1988, 1991) argued that connectionist systems could display systematic behaviour without employing representations with constituent structure. This is often called 'functional compositionality', a term I will use when the behaviour of a system, described in representational terms, displays systematicity in the absence of constituent structure to explain it. Some connectionist networks may well be functionally compositional, especially if they are trained in ways that encourage it.[6] How is such systematic behaviour to be explained? That is the question pressed on Smolensky by Fodor & Pylyshyn (1988). Fodor & Pylyshyn argue that the only way that such performance can be explained is by the operation of compositional structure.

Connectionists have three ways of responding to Fodor & Pylyshyn's challenge. First, they can accept the point but doubt the extent to which cognitive phenomena display systematicity. They can argue that the evidence for systematicity is weak in respect of the cognitive capacities of animals, and in respect of the many aspects of human thought which do not involve language. All of these could be implemented in non-systematic connectionist systems. It is beyond the scope of the present chapter to examine that large body of empirical evidence. I raise it here as an open possibility. Second, if there is evidence that some non-linguistic skills are systematic, then the connectionist can rely upon functional compositionality. These skills could be implemented in connectionist systems which happen to display systematicity. Such systematicity might have been encouraged by the training regime. But, according to this response, the connectionist accepts that there is no deeper explanation, in terms of the architecture of the system, as to why such systematicity arises.

The third tactic is to argue for an explanation of systematicity which does not rely upon compositional structure. Smolensky (1995) attempts to do so. His harmony network is clearly not just a local connectionist implementation of symbolic processing (one of Fodor & Pylyshyn's worries). However, the fundamental objection was not that the representations were local,[7] but that distributed representations could still contain constituent structure. And Smolensky ensures that his representations do have constituent

---

6 Smolensky (1988, 1991); e.g., in a dynamic system, Pollack (1990).

7 In some 'semantic networks' the representations are local, since each node represents one fact or feature. So one worry was that some 'networks' might just be symbol systems with local representations.

structure, by building them out of orthogonal vectors. Since the vectors do not in general align with the dimensions of their component units, a combination of orthogonal vectors will be a genuine superposition – each unit plays a role in representing many of the component vectors at the same time. So there are no local symbols, and the constituent structure does not map onto the network's architecture. But the compositional structure is there, nonetheless. It is what enables Smolensky to employ extraction matrices to arrive at representational components. So, I agree that Smolensky's networks display systematicity. But they do so by building it in: it is a feature of the way in which the network is taken to be representing at all. Thus, Smolensky can explain the systematicity of his networks; but the explanation proceeds in the standard way, via representations with compositional structure, albeit in Smolensky's case implemented in a constituent structure which is non-standard, making it interesting and original in its own right.

Perhaps there is a way to explain functional compositionality without adverting to constituent structure. But as yet, no such explanation has been convincing. So I will accept that, for the range of cognitive phenomena which are systematic (which may be narrower than sometimes claimed), the current options are to explain the systematicity by positing representations with constituent structure, or to leave it unexplained. It may turn out that training can produce functional compositionality for which there really is no further explanation, but that seems unlikely, so it is unattractive to adopt that position until alternative explanations have been extensively investigated. The pressure from Fodor & Lepore's challenge is always to provide an explanation of systematicity which goes far enough to be satisfying, but without going too far – otherwise it amounts to an explanation of why the system under consideration is actually classically computational.

3.2 Compositionality Amongst Clusters in State Space

The last subsection accepted that the best explanation of systematicity is the existence of representations with combinatorial structure. In this subsection I will explain how my theory of connectionist content may allow for a different kind of compositionality, underlying a limited kind of concatenative productivity.

Recall that some clusters in a state space may be superpositions of two or more other clusters.[8] For example, where clusters are found around vectors $\underline{x}$ and $\underline{y}$, and also

[8] Ch. 2, ss. 5.2.

around the vector $\underline{z}$ = ($\underline{x}$ + $\underline{y}$), then there is no need to admit $\underline{z}$ as a semantic dimension. Although activation is found at $\underline{z}$ in response to some inputs, that activation can be accounted for in terms of semantic dimensions $\underline{x}$ and $\underline{y}$. Suppose their contents are P(s) and Q(s) respectively, where s is the presented sample. Then the content of points falling in the cluster at $\underline{z}$ will be (P&Q)(s). The basic semantic dimensions of a state space are those clusters which account for all the others. Principal components analysis is one way of looking for these basic clusters. Similarly in dynamic networks: the component processes that are the semantic dimensions may sometimes superpose.

But now this gives rise to a certain limited kind of productivity: the network can represent P(s) and Q(s), and as a result it can represent (P&Q)(s). The range of productivity is limited. There is no reason to think that arbitrary combinations of semantic dimensions will correspond to further semantic dimensions. Their sums may not lie within the state space at all. Or they may not be reachable from any possible pattern of input. However, where the concatenative regions are within the limits of the architecture, the ability to activate the basic semantic dimensions will entail the ability to activate their concatenation. Isn't this the kind of phenomenon that was supposed to require compositionality?

I make two observations. First, this is only a limited phenomenon. It is not systematicity, since there are no conceptual constituents, nor is it full productivity, since it is very limited by the particular set-up of a network. All that can be said is that in some networks some of the clusters can sometimes be related in ways that reflect concatenation of their contents. My second observation is that the phenomenon still relies upon constituent structure. There is not the kind of general compositionality found in a language of thought, but the representation at $\underline{z}$ which has the content (P&Q)(s) *can* be divided into the components $\underline{x}$ and $\underline{y}$. Those components are superposed rather than classically combined, but they are both present nonetheless. That makes it rather like the superposed vectors in Smolensky's harmony networks (although there the components were necessarily orthogonal).[9]

Does this mean that I have fallen into Fodor & Pylyshyn's trap of providing a connectionist implementation of a classical computational system? Fodor & Lepore (1999) press this question against Churchland's (1998) state space semantics. They ask whether he is describing a processing level or a syntactic level. The former, they say, has nothing

[9] Smolensky (1991), (1995).

to do with content. And, they argue, if the description is genuinely syntactic, then it must be compositional, so it will demonstrate that the connectionist network is implementing a classical computational system. The answer is that my clusters are at both the processing level and the syntactic level. I deny that a distinction is to be made between them. The syntax is found at the processing level. Amongst the various ways of describing the processing mechanism, the syntax is the one that divides the processing into items which are vehicles of content (this idea is explained further in chapter 5 below). However, this does not entail that the system is classically computational, since the syntax is not compositional. As I argued in section (2) of chapter 2, there is no need, minimally, for a syntax to be compositional (although many are).

What of the limited compositionality that sometimes arises when some clusters in state space are composed of others? Does that entail that the system is classically computational, at least to some extent? The answer is no, because there is no sense to be made of being partially classical. The basic syntactic items in connectionist systems are not required to compose, although in some systems some syntactic items may be combinable. Combining is constrained by the overall size of the state space, and the nature of the connections between input and hidden layers. There are no such constraints in the classical case. Indeed, classical computation is characterised by the absence of such constraints – it explains systematicity or productivity precisely because there are no such constraints. The patterns of constraint are quite different in the clustered state space of connectionist systems, as will be the patterns of error, and the type of response to damage. Furthermore, the systems are quite different developmentally: syntactic items in state space arise only as a result of learning (see ch. 2, subsection 6.1). For all these reasons, it is clear that the two models are quite different. That being so, to the extent that clustered state spaces show some productivity, that is an advantage of the model, and not a reason to see it as classically computational.

(4) QUASI-FREGEAN SENSE

The contents ascribed to states of connectionist systems, according to the theory in chapter 2, are found at the level of reference, and do not posit any kind of content at Frege's level of sense. Whether that is a problem depends upon what you think of quasi-Fregean sense. There are a range of theoretical positions. At one extreme are those who reject the relevance of reference and hold that content at the level of sense is the only thing that will feature in psychological explanations (Segal 2000). At the other end of the

spectrum are those who deny the existence of anything psychologically real that corresponds to Frege's level of sense, and so exclude sense from psychological explanations (Millikan 2000).

The representations in a connectionist network do not combine into more complex representations. As discussed in subsection 2.1 above, there are no constituent concepts that can be functionally associated to represent new pieces of information. On one conception, quasi-Fregean senses are individuated in terms of beliefs associated with a concept. For example, two people may have concepts H and P respectively, referring to the very same planet, but differing in that H is functionally associated with a morning concept, and P is functionally associated with an evening concept. If senses are individuated by those kind of considerations, then connectionist representations do not have such senses. Clusters lack the kind of associations that underpin the formulation of such classic Frege cases.

So: no concepts → no sense? Not exactly. Because there is another level at which state spaces can be compared. Recall that Laakso & Cottrell's measure is a sufficient test for same content (when applied between systems operating on the same samples), according to my theory, but is too strong, since it takes account of topographic arrangements in state space between contentful points, as well as their contents.[10] Is this a separate contentful level for comparing networks? That is what Tiffany (1999) suggests.[11] It is quite unlike a level of sense individuated in terms of associated beliefs. First, arrangements of clusters in state space do not reflect or represent any fact – they are not interpreted as the network making any claims. Contrast, for example, the fact that *Hesperus rises in the morning*, associated with the concept HESPERUS. To see the difference, observe that to think *Hesperus rises in the morning* a person has to activate both his HESPERUS concept and his MORNING concept. By contrast, when a network produces activation in a cluster, it does not activate any other clusters. Second, in a connectionist system there is no parallel to the process of making new conceptual connections. Clusters differentiate during development, but they can't be further functionally connected to encode some additional fact. Of course, as I have observed, clusters may well play a role in a wider conceptual system.[12] But such associations do not arise within a single state

[10] Ch. 2, ss. 4.2.

[11] Ch. 2, ss. 8.2.

[12] Ch. 2. ss. 7.2.

space. So the topographic arrangement of clusters does not look like another contentful level. Is it important, nevertheless?

Calvo Garzón (2003) argues that networks trained to classify the same properties, but encountering different sets of samples bearing those properties, will differ at this level of topographic arrangement. He uses that claim to reject state space semantics. My theory avoids this objection since such topographic differences are not reflected in the contents ascribed to clusters (as explained in chapter 2, subsection 4.2). Could it then be seen as a virtue that different networks, with different experiences of objects in the same domain do have different topographical arrangements of contentful clusters? Calvo Garzón draws an analogy with the different ways in which a dog breeder and a non-owner would think of dogs. He takes different topographic arrangements to reflect these kind of differences.

My position on this level of comparison is as follows. First, it is not quasi-Fregean sense. Second, whether it is important depends upon further empirical work. If similarity and difference in such topographical arrangements is empirically useful, then there is no problem with admitting its existence. Alternatively, it may be a way of comparing networks which has no utility in connectionist modelling and no applicability to biological systems. That is an open question. The first indications, from Laakso & Cottrell's work, are that it may be a useful level of comparison. If it is important, then Laakso & Cottrell's measure tests for similarities at this level directly.

Laakso & Cottrell's test is just one of many ways of measuring whether two networks trained to perform the same task have hidden layer state spaces with the same geometry. From a theoretical point of view the details of their measure – using rank orderings of inter-activation distances and the GPA test – are unimportant. Any method of testing whether two layers have the same clusters in the same topographical arrangements will be equally valid.

Just one example of an alternative is the method employed by Goldstone & Rogosky (2002). They use their algorithm to compare the metrical similarity spaces of two conceptual systems, but it would be equally applicable to the task of comparing the arrangement of clusters in two different networks. It uses a constraint satisfaction network to arrive at what they call 'correspondences' between concepts.

Does similarity at the level of topographic arrangements succumb to Fodor & Lepore's charge of holism?[13] It need not, because the clusters themselves are not individuated holistically. We can take a cluster and assess which other clusters are nearby and far away in state space, since the identity of a cluster is not determined by its position with respect to other clusters, but by which samples produce activation within it. The comparison takes clusters individuated contentfully at the level of reference, and asks what the relations are between them in state space. So there is no threat of content holism. Moving one contentful point will alter the topographic arrangement, but will not change the content of any of the points. Granted, it may be rare for networks to have exactly the same topographic distribution of contentful clusters (although Laakso & Cottrell's work shows they can be very similar). But that is not holism. Holism arises if the conceptual interrelations form part of the content of each concept, so that a change in any one interrelation ramifies, and entails a change in the content of all concepts in the system.

As an aside, a similar move can be made in individuating quasi-Fregean sense in a conceptual system. Those who advocate two levels of content, both sense and reference, can insist that conceptual content is determined only by interrelations amongst referential contents, and not by interrelations amongst conceptual contents. Say a thinker believes that *dogs are hairy* and that *cats are hairy*. Now consider the relationships thereby established between the concepts DOGS, CATS and HAIRY. The sense of DOGS then depends upon the referential content of HAIRY, and the sense of HAIRY depends in turn upon the referential content of CATS. However, the sense of DOGS does not depend upon anything about the concept CATS, since it does not depend upon the sense of HAIRY. This move is not widely deployed, however, since most philosophers believe that sense determines reference. That, certainly, is Frege's idea. In which case, the sense of DOGS would depend upon the sense of HAIRY, since it depends upon the reference of HAIRY, which is determined by its sense. And so the sense of DOGS would depend in turn upon the sense of CATS, and of every other concept with which it is interconnected. Holism does not follow directly from the existence of a separate level of sense, distinct from referential contents, but rather from the idea that sense determines reference.

[13] Fodor & Lepore (1992), (1993) & (1999).

(5) DIFFERENTIATION INTO BELIEFS AND DESIRES

Representations may represent how things are, or what the system should do. Millikan argues that the most primitive representations carry both sorts of content simultaneously. They are what she calls 'pushmi-pullyu' representations (Millikan 1996b). For example, the slap of a beaver's tail means both *danger here now* and *dive to safety*. The same token carries both kinds of content at once. It is only in a representational system that differentiates beliefs from desires that the two kinds of content are carried by different vehicles.

The connectionist systems of chapter 2 do not differentiate into beliefs and desires. There are not different types of cluster specialised for either the input or the output role. However, recall that the attribution of content is sensitive to both input factors (properties of the samples) and output factors (relevance to the output classification).[14] If Millikan is right, then the clusters are pushmi-pullyu representations. Their content should then be something like: *the currently-presented sample has property P and act at the output layer so as to produce a P-appropriate classification*. Whether the content of representations should 'look both ways' like this is an issue examined further in chapter 6 below (sections (8) and (9)). The theory of content in chapter 2 is conducive to that approach, and even suggestive of it to the extent that it relies both on a system's inputs and outputs in determining content. One caveat should be discussed here, while we are considering the extent to which the connectionist approach can be extended to some human representational systems. That is whether representations which 'look both ways' are necessarily less determinate than those found in systems which do differentiate into beliefs and desires.

Papineau (1993) takes the view that only in a system that differentiates between beliefs and desires can representational content be ascribed with adequate determinacy.[15] Millikan does not see any particular determinacy problems with the content of pushmi-pullyu representations. Nevertheless, she too thinks that a belief-desire psychology works towards determinacy; but that is because representations are consumed by different

---

[14] Ch. 2, ss. 6.3.

[15] Papineau (1993). He has now expanded that view to allow for the existence of content relative to a consuming system; so that in more primitive systems the representations may not have determinate contents *tout court*, but may have adequately determinate content relative to the function of some consuming system, which must be independently specified (personal communication).

systems for a number of different purposes in such cases, so content is triangulated amongst the different uses, making it more determinate (Millikan, manuscript).

My connectionist contents do not seem to suffer from damaging indeterminacy, even if they do have 'look both ways' contents. But that may be because they rely upon the determinacy of the output contents. Indeterminacy there would infect the content of hidden layer clusters. Thus, it is important for my theory that there is some way of ascribing reasonably determinate output contents to realistic analogues of my systems.[16] That should be borne in mind when we discuss biological systems in the next section. The possibility should be kept open that the theory in chapter 2 imports a degree of determinacy which, although appropriate when explaining the connectionist systems created by modellers, is unwarranted when a similar approach is extended to natural representational systems.

My answer, in short, is that it does seem that connectionist-type systems in biological brains are embedded in sufficiently determinate representational contexts that their hidden layer clusters will have adequately determinate contents to make contentful explanation useful. However, since the connection between models and real brains is still not very strong, exactly how this issue turns out will depend upon the results of further empirical work.

(6) Real Brains and Unsupervised Learning

6.1 Distributed Representation in the Brain

So far in this chapter we have seen that the approach to connectionist content that I have been exploring does not treat of concepts, shows only very limited compositionality, and does not admit of contents at the quasi-Fregean level of sense. Those considerations importantly limit the extent to which the approach can be applied to human cognition. Nevertheless, there are many aspects of human and animal cognition which lack those sophisticated features, but which are still representational. Thus, the possibility remains that my connectionist approach can be extended to some such cases. The present section suggests how.

The first task is to show that some systems in real brains do indeed employ distributed representations. In connectionist systems, a vehicle of content consists of

[16] Cf. ch. 2, ss. 3.6.

activation across a whole layer of nodes. The analogue in brains is activity across a whole population of neurons. The converse is local representation. A representation in the brain is localist where the syntactic item consists of the activity of a single neuron. The mythical 'grandmother neuron' would be an example: a neuron whose function is to represent a person's grandmother, which fires when and only when the person experiences his grandmother, and which is employed in downstream processing in ways that are specifically relevant to grandmothers.

Experimental practice tends to focus on local representations, because finding them is a lot more tractable. Electrophysiological studies use electrodes implanted in the brain of a live animal to record activity in a single neuron.[17] Experiments can show how the activity produced depends on the types of stimuli presented to the animal or the types of task it is asked to perform. It is found that some such 'single-units' have very precisely delimited sensitivities. The temptation is to conclude that their job is to represent the category with which they correlate. This is rather like the connectionist modellers who do scatter plots of individual nodes in order to find out what they represent. The method is flawed because the unit recorded could be part of a cluster which represents in a distributed fashion. Imaging studies similarly look for areas that light up preferentially for certain classes of stimuli. Experimenters sometimes conclude that such an area represents the class of stimulus which activates it. However, these areas are differentiated with much lower resolution than single-unit recording, being based on the blood-flow response in an area containing thousands of neurons. So here it is even more likely that representations of a particular domain (e.g., faces) are distributed across the area, with different faces represented by different distributed patterns of activity. The voxels light up preferentially to faces on an fMRI scan, for example, because each representation of a face leads to a (different) pattern of activation throughout that area.

Downing et al (2001) is one example of the use of imaging to support the localist approach. Building on work to identify brain areas which respond selectively to spatial layout (the parahippocampal place area) and to faces (the fusiform face area), the authors identify a region of the lateral occipital cortex in the human brain that responds

[17] In practice, most electrophysiological studies can only plausibly claim that their electrodes measure activity in a small block of neurons. For this reason, the technique is often called 'single unit' instead of 'single cell' recording. However, this is a limitation of the experimental technique which is being gradually overcome. The aim is certainly to record activity in single neurons.

selectively to pictures of the human body. They call this the extrastriate body area, and claim that it represents human bodies.

But even imaging can support the opposite conclusion. A rare example of imaging work uncovering distributed representation is Haxby et al (2001). They use an fMRI study to argue that the brain represents categories by means of distributed patterns of neural firing. Their results agree with many earlier studies – that certain brain areas are more responsive to such specific categories as: faces, cats and houses. But they also found that global patterns of activity in the ventral temporal cortex *excluding* the maximally responsive area could be used to predict the category of stimulus presented. Furthermore, even taking just a region maximally responsive to one category of stimulus (faces, say), the pattern of activity within that region could predict which category other stimuli (such as cats) belong to. Haxby et al conclude that representations of items in each of these categories are distributed across the ventral temporal cortex. Martin agrees that object concepts are represented by distributed clusters in the ventral temporal cortex.[18]

An overriding problem is that the experiments do not arbitrate conclusively between the hypotheses of localised and distributed representation. Haxby et al's results can be interpreted in a number of ways which are consistent with local representation. For example, the patterns in other brain areas may be caused by a local representation of the given category (for example, being an appropriate response to it); they may be part of the common processing leading up to local representation of objects in that category; or they might be an incidental response of other areas, deriving from similarities in the stimulus, but unrelated to how the category is represented. Similarly in the other direction, Downing et al's findings are consistent with the hypothesis of distributed representation – specific brain areas being more active in representations of a given type, but a large number of neurons, both active and inactive, constituting any given representation. Behavioural double dissociations associated with selective brain lesions are sometimes taken to suggest local representation. However, there is usually a behavioural deficit in respect of whole category of stimulus. So again, the data only show that items from the affected category are represented within the damaged area: they could either be represented locally within the area, or distributed across it. Thus, neuropsychological findings are consistent with the distributed hypothesis.

---

[18] Ishai, Ungerlieder, Martin et al (1999), Martin (2002).

One might begin to suspect that the dispute between local and distributed representation concerns a distinction without a difference. However, the connectionist model shows that there is a distinction to be drawn. In a connectionist system, what happens at the output layer typically depends upon what is happening in each of the nodes of a hidden layer. Each is contributing to downstream processing, and ultimately to performance of the function which the connectionist system is required to execute. In the same way, a real distinction can be made between local and distributed representations in brains – by looking at whether the way some brain system performs its function depends upon the simultaneous activity of a number of neurons. Of course, to apply this distinction in practice requires scientists to identify the functional organisation of the brain in information-processing terms. And it is hard to do that without knowing what the representations are (to say the least). However, as understanding of what is going on in the brain increases it will be easier to draw the distinction between local and distributed representation in practice. The results so far at least leave open, and often favour, the distributed approach in many domains.

Further support for the distributed hypothesis derives from the successes of connectionist modelling. Moving to distributed representations has been a profitable strategy for modelling a wide range of tasks. That raises the possibility that there are tasks which humans perform that can only realistically be achieved by computing over distributed representations. A more cautious claim is that there are many tasks for which a solution using distributed representations was more accessible in the adaptive landscape than one employing local representations, being an easier way to do the task with the sorts of resources available in the brain. If so, since humans have evolved by natural selection, distributed mechanisms are more likely in such cases.

In what ways are distributed solutions more efficient than localist ones, given the types of resources available for information processing in the brain? I will mention two. Firstly, more things can be represented with a given number of neurons. Even if limited to binary encoding, the number of positions in representational space is an exponential function of the number of neurons in a distributed representational system, but only a linear function of the number of neurons if each represents locally (Rolls & Treves 1998, p. 13). Secondly, the speed of processing is much faster with distributed representations. In a real system with some background noise, and where neurons take time to build up to firing, a system reading local representations would have to wait for several neuronal firings before being able to tell which of a population of neurons is firing at an elevated

rate. The information in a distributed representation can be extracted much more quickly. The very first firings provide a rough indication of which area of state space the representation falls in, even if a few are due to noise, so there is no need to wait for the time taken for several firings before acting on the representation in downstream processing. Just how much of a difference this makes depends upon the exact model used: how the neurons respond dynamically and whether they remain close to the firing threshold (Rolls & Treves 1998, pp. 265-266). However, it is another indication of possible benefits of the distributed approach.

In sum, there are some good reasons to think that some representations in real brains are realised in a distributed fashion by means of the firing of a whole group of neurons. The theory in the last chapter relied not only upon distributed representations in connectionist systems, but also on the existence of clusters in state space. The state space of a brain area has a neural dimension corresponding to each neuron in the area.[19] It is to be expected that a number of different sensory inputs will produce similar distributed representations. If so, those inputs form a cluster in state space. Thus, to the extent that brains do employ distributed representation, clustering is to be expected. The remaining question is whether any behavioural results should be explained in terms of processing over such clusters and, if so, what contents should be ascribed to them. That is the topic of the next subsection.

6.2 Processing Over Clusters

The theory in chapter 2 was based on feedforward networks trained by backpropagation. It is not known whether anything like this occurs in real brains. However, other connectionist models are much more neurally plausible. In this subsection, I will argue that these models, too, should be viewed as processing over clusters.

To explain why these connectionist models are biologically plausible, I start by setting out some empirical findings on the mechanisms that implement classical conditioning. The mechanisms rely upon Hebbian synaptic plasticity – the idea that coincidence of activity pre- and post-synapse leads to a stronger connection between the pre-synaptic neuron and the post-synaptic neuron. By giving two examples, I will show that the molecular-level implementation varies. The examples are gill withdrawal in the

[19] Recall that neural dimensions do not correspond to semantic dimensions. The latter depend upon where in state space clusters fall.

marine snail *Aplysia*, and the conditioning of the mammalian eye-blink response. They are standard cases where classical conditioning at the behavioural level has been explained in terms of implementing mechanisms. Neither case involves distributed representations. Indeed, it is not obvious that, in these examples, the behaviour need be described in representational terms at all. Furthermore, if the internal processes are attributed contents, then the representations are clearly localist, not distributed. We have seen above that many brain systems employ distributed representations. How is classical conditioning in such systems to be explained? Connectionist models show that the same basic form of synaptic plasticity – Hebbian learning – can also account for classical conditioning in systems which *are* representational, and where the representations are distributed. Thus, the value of the two examples is to show that Hebbian neural plasticity exists, and is realised in a variety of ways. Connectionist models then show how that mechanism can give rise to processing over distributed representations. I will argue below that these models should be understood in terms of clusters – conditioning arises because of associations formed between clusters. Thus, the approach to syntax advocated in the previous chapter should be adopted: processing should be understood as occurring over clusters in state space.

The other type of associative learning is instrumental conditioning, which has also been extensively studied and described in a range of animals and experimental situations. However, it is less clear how instrumental conditioning is implemented neurally. The connectionist models described in the last chapter seem, behaviourally, to be examples of instrumental conditioning. But there is no known analogue to the process of weight adjustment by backpropagation of error. That algorithm relies upon calculating global quantities – gradients in error space – and using them to adjust all of the weights. It is at best controversial whether that sort of thing occurs in brains. Nevertheless, instrumental conditioning is clearly widespread, so it is implemented somehow; most likely, in a variety of ways. Given the importance of distributed representations in the brain, implementations of instrumental conditioning may well employ distributed representations. What I argue below is that, when instrumental conditioning is implemented in a system that employs distributed representations, it should be explained in terms of processing over clusters. Furthermore, instrumental conditioning could sometimes depend upon the development of new clusters, as a variety of sensory inputs are associated into a new cluster because of their common relevance to some output task.

Thus, this subsection will establish that there are strong empirical grounds for thinking that some real biological systems will be best understood as processing over syntactic items which are clusters of distributed representations in state space. Contents can then be ascribed along the lines of the theory in the last chapter.

I start with classical conditioning (also called Pavlovian conditioning). In classical conditioning, an animal begins by responding in an appropriate way to a biologically significant stimulus (the unconditioned stimulus, "UC"). For example, a dog salivates in response to food. A neutral stimulus (the conditioned stimulus, "CS"), such as the sound of a bell, is then consistently presented with the UC. As a result, the animal learns to produce the same response to the new stimulus CS, even in the absence of the original one UC. Thus, after conditioning Pavlov's dogs salivate on hearing a bell.

In the marine snail *Aplysia* classical conditioning of gill withdrawal is understood right down to the molecular level. Strong stimulus of the sensory nerves from the tail (UC) elicits gill withdrawal; weak stimulus from the mantle (CS) does not. However, if the CS is followed by the UC, after a few trials the CS comes to cause the gill withdrawal on its own. The mechanisms responsible have been extensively investigated (Kandel and Hawkins 1992). Neuronal pathways from the UC and CS converge on the motor neuron driving the gill withdrawal. This motor neuron is able to act as a coincidence detector: when stimulation from the UC immediately follows the CS, molecular changes ensue sensitising the synapse from the CS. These pathways are described at the molecular level: only coincident activity is sufficient to amplify a crucial enzyme, the heightened activity of which makes the synapse between CS and response more sensitive. In short, in *Aplysia* a molecular mechanism has been discovered that gives rise to Hebbian learning: coincident activity leading to stronger connections.

A second example shows how a similar result is achieved by a different mechanism. The mammalian eye-blink reflex can be classically conditioned by playing an audible tone (CS) just before a puff of air is delivered to the eye (UC). The circuits responsible for this reflex and the conditioned change are low-level, found in the cerebellum (Shepherd 1994, p. 630). Again, pairing of the CS with the UC leads to molecular changes, the result of which is that the CS comes to elicit the eye-blink in the absence of the UC. The molecular mechanism is different, and proceeds by long-term depression of a key synapse (depression rather than potentiation can do the job since it is a synapse to an inhibitory neuron). Again, synaptic modification proceeds via coincidence detection.

Studied at the level of animal behaviour, classical conditioning is not a unitary phenomenon. It varies between different animals and depends upon the experimental situation in which they are trained.[20] For example, some kinds of response depend mostly on the nature of the UC, others more on properties of the CS. By giving two classic examples of implementing mechanism, I have illustrated that classical conditioning is implemented in different ways at the cellular level, but that coincidence detection is a central feature. That aspect is modelled by the Hebbian synapses employed in many unsupervised connectionist networks.

Although the localist cases are most easily uncovered experimentally, the same type of plasticity can occur between distributed representations. Hebbian strengthening at each of the synapses between two distributed representations would increase the association between them. However, a single episode of coincident activity is not sufficient to develop an association. This is where clusters come in. Where the activity produced by different inputs falls into clusters, associations can form between those clusters. The effect of clustering will be that successive CS inputs will produce activity in the same region of state space. Similarly for successive UC inputs. Each pair of stimuli will gradually strengthen associative connections at the synapses active between the two clusters. The result will be an association between the two clusters. That distributed association can arise from the action of purely local Hebbian plasticity. I will explain below how connectionist systems can model this process. Even if the representational explanation seems dispensable for *Aplysia*, and perhaps for the eye-blink response, it is not so easily dismissed for systems that are more complex and distributed. Thus, we arrive at plausible models where Hebbian learning gives rise to processing over clusters – the formation of new associations between pre-existing clusters.

The simplest connectionist model of classical conditioning is the pattern associator network with a single output neuron (Rolls & Treves 1998, pp. 16, 23-41). The network has two inputs: a local UC, carried by a single wire, and a distributed CS, consisting of patterns of activation across a layer. Each input from the CS connects to the output. When the output is activated by the UC, connections are strengthened between it and those of the CS inputs which are active at the same time. This is achieved by local Hebbian plasticity. When successive patterns of distributed CS activity fall in the same cluster, an association will gradually develop between activity within that cluster and the UC. Thus, activity in

[20] Pearce 1997, p. 51 and ch. 2 *passim*.

the cluster will eventually be sufficient to give rise to the output response on its own, without the UC. That is to describe the mechanism in terms of processing over clusters: the CS is a cluster, which gradually becomes associated with the UC, leading to classical conditioning. This connectionist model plausibly captures the mechanism giving rise to some examples of conditioning in real biological systems. For example, the UC may be a particular taste, which is represented in a single neuron with a sensitivity specific to that olfactory input. The CS could be a distributed pattern of activity driven by visual input. CS clusters would consist of visually similar inputs. The pattern associator network shows how an organism that already responds in a certain way to that UC taste could be classically conditioned to respond in the same way to samples on the basis of their visual appearance (even when that visual appearance is represented distributively). Indeed, classical conditioning between visual and olfactory stimuli may well be implemented in exactly that way in the amygdala (Rolls & Treves 1998, p. 150).

The single output pattern associator network does not model the development of new clusters. It does show, however, that classical conditioning may arise, in real organisms, because of processing over clusters, implemented by a local Hebbian mechanism.

The model generalises to cases where the UC is also distributed. These are general pattern association networks (McLeod, Plunkett & Rolls 1998, ch. 3; Rolls & Treves 1998, ch. 2). Before training, one layer of nodes drives the output (the UC). Each of these is connected to each of the nodes from a CS. The connections are Hebbian. Thus, if distributed activity from the CS consistently falls in one region, and distributed activity from the UC consistently falls in another (each in their own respective state spaces), then temporal coincidence will lead an association to develop between the two clusters. Eventually, the CS cluster will drive the outputs, even when the UC cluster is not activated endogenously. The processing that implements this classical conditioning occurs between clusters in state space.

Competitive networks also act on clusters. They transform distributed activity within a cluster into a single localist output. (The idea of converting distributed clusters into discrete outputs was mentioned in subsection 7.2 of chapter 2.) In a competitive network, each input is connected to all of the outputs. The outputs compete to determine which has the largest response. That encourages single unit outputs. Where input activity falls into distinct clusters, each cluster will gradually become associated with its own output. As a result, clusters that already exist in some state space can be converted into

unitary outputs. Thus, were clusters to develop in some intermediate layer, as they do in the connectionist models of the previous chapter, these clusters could be converted into unitary outputs by the operation of competitive networks in the brain, facilitating the use of these intermediate clusters for new purposes. Again, the biological systems modelled by such networks should be understood in terms of processing over clusters.

Autoassociation networks perform another operation on clusters. These networks are not Hebbian, but are thought to be neurally plausible (McLeod, Plunkett & Rolls 1998, ch. 4; Rolls & Treves 1998, ch. 3). An array of external inputs drive an output array directly (without cross-connections), but also connect back so that each output acts as an internal input to every node in the array. The aim of the network is to filter out noise, and to continue producing at output the pattern presented at input. Weights are adjusted by a 'delta rule' algorithm, which acts at a connection so as to decrease the difference between internal and external inputs. The effect of training is to form associations between external and internal inputs. Since these are both driven by the same patterns, the result is to draw together a cluster of inputs into a single vector output falling within that cluster. From the point of view of clusters, the autoassociation network acts as a focusing lens, making the cluster tighter. Again, the network's operation can be clearly understood in terms of processing over clusters.

Instrumental conditioning is the other paradigm of associationist learning. It occurs when reward or punishment modifies an animal's behaviour. Can connectionist networks model the mechanisms that give rise to instrumental conditioning? The answer would be straightforwardly affirmative, following from the discussion above, were instrumental conditioning to be just a special case of classical conditioning, as some have claimed. Consider, for example, a rat that receives a food pellet on pressing a lever. The rat will learn to press the lever to obtain food. Is this just a classically conditioned association between the sight of the lever and the arrival of food? That kind of explanation has been empirically excluded by experiments that rely upon bi-directional control (Grindley 1932). These studies show that animals can be instrumentally conditioned not only to produce a response which might have been classically conditioned, but also, within the same experimental set-up, the opposite response (e.g., turning the head in the opposite direction in order to receive food) – which could not have been classically conditioned.[21] Thus, instrumental conditioning is a different phenomenon, and

[21] There are also theoretical attempts to assimilate classical conditioning to instrumental conditioning, which

calls for a separate explanation. Instrumental conditioning is found in a very wide range of experimental situations, and in a wide variety of animal species (Pearce 1997, ch. 4). No simple explanation of the mechanism of learning will cover all these cases. An account of the varying influence of this variety of factors in different situations will rely upon a number of different explanations, or some rather complex unifying model (Dickinson 1994). Correlatively, if connectionist systems are to model instrumental conditioning, the model will vary from case to case.

To the extent that anything general can be claimed, instrumental conditioning looks more like the kind of supervised learning which was the focus of chapter 2. The existence of conditioning based on reward and punishment, even in relatively simple systems, shows that there must be mechanisms of neural plasticity that rely upon an error signal, or some target output against which training can take place. Instrumental conditioning allows an animal to become better at generating the sorts of actions that will produce beneficial results. However implemented, it results in new connections between perceptual experience and action. Where distributed representations are involved, similar perceptual experiences will fall into clusters. Thus, the result of instrumental conditioning will be to associate existing clusters with new outputs. For example, the rat learns a new range of situations in which to approach and press a lever. The mechanisms of these changes are not well understood, but the important point is that they seem to take place over clusters. The effect of reward and punishment may be to form new associations between existing clusters. However, it may also give rise to the formation of new clusters, as in the systems described in the previous chapter.

To summarise, we have several models of how operations are carried out on clusters: association, focusing and conversion into unitary outputs. What should the theorist say about syntactic development and content ascription in these models? No single story covers all the examples. There is a spectrum of possible operations:-

(i) The formation of new connections based on pre-existing clusters, without any syntactic development or change in the content to be ascribed to the clusters.

(ii) The association of an existing cluster with new inputs or outputs, not resulting in new clusters, but leading to a change in the content to be

---

are equally controversial (Pearce 1997, pp. 100-101).

ascribed to the existing syntactic item, because of its new input sensitivities, or the new ways it is used.

(iii) The formation of new clusters out of existing clusters, and thus the formation of new syntactic items to which new contents should be ascribed.

(iv) The formation of entirely new clusters, to which new contents should be ascribed.

Competitive networks and autoassociation networks are examples of (i). Processing occurs over clusters, but there is no development of entirely new clusters, nor is there any reason why the content to be ascribed to clusters should change. These networks act so as to make clusters more available to downstream processing, by focusing them or converting them into localist representations. The pattern association network with a single output may also be of type (i). A pre-existing cluster of activation patterns in CS inputs is associated with a new behavioural output. The content of the cluster need not have changed as a result. On the other hand, the new output purpose for which the existing cluster is employed may alter the content which should be ascribed to it – type (ii). In either case, the model does not account for syntactic development. A general pattern association network may, however, give rise to syntactic development, falling under case (iii) above. That model shows how an association can develop between clusters from UC and CS inputs. As a result, a wider range of sensory inputs will give rise to activation within the same region of the state space of the neurons which originally responded only to the UC stimuli. Consider that state space after training. A whole new range of sensory inputs, which before did not produce activation in that state space at all, now fall within it. That is, clusters can be discerned in that state space in respect of a much wider class of input samples. The samples in the new clusters result from merging the samples in the pre-existing clusters between which associations formed. Since these are new clusters, it follows that new contents should be ascribed to them, to reflect the new range of samples falling within the cluster.

Finally, instrumental conditioning may provide examples of type (iv). This will occur, for example, if instrumental conditioning leads to the creation of entirely new intermediate clusters, as found in the connectionist systems of the previous chapter. Rewards and punishments may be biologically-salient purposes against which the development of new clusters takes place. If so, new clusters would be identified by

considering the distribution in neural state space of the new samples which, as a result of learning, come to perform the function which drove the conditioning process. For example, where conditioning is driven by food rewards, the clusters are individuated by considering the distribution in neural state space of samples of the foods on which the animal was conditioned. Furthermore, the purpose served by the outputs on which instrumental conditioning occurred play a role in individuating their content: the clusters should be ascribed contents relevant to acting in food-appropriate ways (cf. ch. 2, ss. 3.5.1).

That gives an indication of how the clustering proposal can be extended to processes that take place over distributed representations in biological systems. Further development will depend upon improvements in understanding the processes that occur in real brains. As a result, the account given here is necessarily tentative. What should be clear, however, is that clusters provide the syntax of such models. All the signs are that, when the manipulation of distributed representations by real brains is fully described, it will be understood in terms of processing over clusters.

(7) Criteria for Extending the General Approach in Chapter 2

I have argued that my theory of content in chapter 2 can be extended to other types of connectionist systems, that are plausible models of some biological systems; and that my theory can extend to the mechanisms of classical and instrumental conditioning in such real biological systems. Does the approach generalise any further? In this section, I will summarise the basic features of the approach in chapter 2. Where these are found in other domains there is a good prospect of bringing the same kind of theoretical approach to bear. The four key features of the theoretical framework of chapter 2 are as follows.

First, the representational states under consideration are intermediate between inputs to and outputs from the system. A range of physically different token intermediate states are available, such that different inputs can produce different intermediates, which can in turn give rise to different outputs.

Second, there is some representational development. New representational types arise as a result of training to meet some goal or perform some action. So the system is seen as acting in some context, and this provides both the causal drive for the development of the syntactic types, and part of the basis for the ascription of content.

Third, intermediate states of the system are subject to a similarity metric. This allows the responses to different inputs to be considered within a similarity space. The

similarity measure is relevant to processing within the system —it is similarity from the point of view of downstream processing. Thus, similarity is an intrinsic property of the entire system.

Finally, learning alters this similarity metric. The result of learning is that different input samples lead to token intermediate states which produce similar results in downstream processing. Learning is aimed at some output task, and proceeds by altering which token states are similar with respect to downstream processing relevant to that output task. Learning draws together token states on the basis of shared relevance for the output task.

In such cases, the result of learning will be to cluster together token internal states, as judged by the similarity metric. If the output task is already understood in contentful terms, then the clusters can be ascribed content based on those input sensitivities which are common to and distinctive of a cluster, and relevant to the output task.

The examples in the previous subsection do not all have these key features. Although all involve processing over representations, some involve only the formation of new connections between existing representations, without alteration of the content of those representations. But these features provide a framework for thinking about different kinds of cases where representational development does plausibly occur.

For example, my framework might help with Barsalou's theory of concepts. Barsalou argues that conceptual thought consists in the manipulation of perceptual symbols (Barsalou 1999). Each tokening of a concept is a 'situated simulation' in perceptual systems that is relevant to the current context. Barsalou's idea is of diverse perceptual representations being associated together in the same 'simulator' (Barsalou 2003). Such simulators are cross-modal patterns of associativity that treat quite different patterns of activity as similar for downstream processing. Barsalou argues that empirical evidence establishes the existence of such simulators, and that they are clearly representational (Barsalou 1999, 2003). However, he says little about how their content is determined.[22] My approach might help. Simulators can be seen as giving rise to a similarity metric in a very high dimensional cross-modal space. Concepts are clusters in that space. The concepts will then refer to things in the world, since such similarities have built up due to

[22] Barsalou (1999), Commentary at pp. 610-611 (Aydede) and pp. 632-633 (Siebel), Response at p. 638.

similarities in the world which are relevant to output tasks. That can be the basis for attributing content to them.

The emphasis on the output contexts in which clusters differentiate should underline that this is not learning inscribed on any kind of blank slate. The structure of the motor systems and behaviour into which processing feeds, and the pre-existing structure of the spaces of perceptual receptivity, will both strongly constrain the way that clusters differentiate. This is consistent with empirical evidence that many human-relevant categories emerge partly as a result of the ways humans are structured physically. For example, evolutionary modelling indicates that phonemes differentiate as they do because of dynamic interactions between the physical construction of the human vocal tract and the way that human hearing works.[23] Thus, clusters do not arise out of nowhere. Nor are they just some neutral characterisation of the inputs from which they derive. Rather, the way that state space differentiates into clusters will be heavily influenced by a whole range of factors, peculiar to the particular organism in which the representations are found.

(8) CONCLUSION

In this chapter I have argued that my theory of content in connectionist systems can be extended to apply to some biologically plausible cases. Having set out some limitations, largely based on the absence of conceptual structure, we found that associative learning is a domain where the approach can fruitfully be applied. But didn't associative learning go out with the demise of behaviourism? Didn't the cognitive revolution teach us to give up reliance on such an anaemic explanatory framework? Shouldn't we be explaining human behaviour with more cognitive resources: learning by testing hypotheses, setting parameters, making analogies, and so on? That is an overreaction to behaviourism. This chapter shows why. Associationist learning mechanisms should not be jettisoned at the same time as we boot out the behaviourist.

The objectionable part of behaviourism was its rejection of internal representation. The success of cognitivism demonstrates that cognition must be seen as consisting of the manipulation of internal representations. However, it is a step too far to suppose that such representations must always have conceptual structure, being computed

[23] Steels (1996), Steels & De Boer (1996).

in a language-like medium of thought. Connectionist models show that very many cognitive-type behaviours can be achieved without any such structure. The burden of this chapter and the last has been to show that we only understand the operation of these networks if we attribute representational content to their internal states. So, a connectionist should not be behaviourist. He should be keen on internal representation. His distinctive contribution lies in showing that, outside language, much of the performance of humans and other animals may not arise from classical-style computation.

Therefore, the connectionist can be seen as taking up the legacy of the old associationists, but enriching it with internal representation, and with all the additional tools that have been brought to connectionist modelling, for example: supervised and unsupervised learning rules, non-linear activation functions, and accounts of the dynamics of networks (attractor and component processes). This gives them the power to explain some important aspects of cognition.

4

Typicality Effects and Prototypes

(1) INTRODUCTION

One of the great success stories of experimental psychology has been the discovery of typicality effects. Typicality is a measure of how like other members of a kind a particular instance of that kind is. The typicality of an object as an instance of one or more kinds will predict and explain many aspects of the way people behave in relation to it. The paradigmatic studies ask subjects quickly to categorise instances: to say what category they fall under. Subjects are separately asked to assess how typical these instances are as members of the category. The primary result is that the speed and accuracy of rapid categorisation judgements correlates with subjects' typicality ratings. Thus, the more typical a stimulus is as an instance of a category, the quicker it will be judged to fall within that category. Similarly, instances that are more typical elicit fewer errors in category judgement. These results have given rise to a large body of empirical work investigating typicality effects. Most theories account for typicality effects by means of prototypes, which are psychologically real structures possessed by subjects.

I suggested in chapter 2 that some typicality effects may be modelled by connectionist systems (subsection 7.3). The current chapter provides an overview of the experimental results, and then explains how connectionist systems may account for some of these effects.

Strikingly, connectionist models can give rise to typicality effects without there being anything like a prototype inside the system. That is radically different from the standard psychological approach, since it dispenses entirely with psychologically real prototypes. Standardly, category judgements are thought to involve checking the encountered instance against lists of prototypical features associated with a concept. The more prototypical the instance, the quicker and more accurate this checking will be.

Thus, conceptual prototype structure is used to account for the typicality effects. In this chapter I argue that prototypes can sometimes be dispensed with. Connectionist models show how some typicality effects can arise without the agent possessing a prototype. I don't claim that connectionist models can model or account for all typicality effects. No doubt, some of the most cognitive results call for the existence of prototype structures that are psychologically real. However, the connectionist model shows that the move from typicality effects to psychologically real prototypes is not always justified, and so must be made carefully.

In short, the job of this chapter is to summarise the empirical data, to explain how it is usually accounted for in terms of conceptual structure, and to argue that some of these results can instead be understood in terms of clusters in the state space of connectionist systems.

The chapter makes another important point. I have argued in several places above that the structure associated with clusters – their topographic arrangement in state space – is not determinative of their reference (instead, it is part of the machinery for individuating the vehicles of content).[1] If that seems rather an odd idea, further support for it can be derived from the discussion of prototype theories below. Many experimental psychologists have used the typicality data as the basis for theories of concepts which individuate concepts in terms of their prototype structure. It is usually assumed, tacitly or explicitly, that the prototype determines the reference of a concept. But that is implausible, for reasons that I explain below. That is, even if a concept does have an associated prototype, the prototype does not determine reference. I arrive at this result independently of any consideration of connectionist models. That is an independent justification for the idea that associated structure need not determine reference. So it supports the claim that the topographic arrangement of state space does not determine content.

Thus, I argue for a distinction between the referential role of concepts and the way they are used to explain subjects' categorisation judgments. In the former role, concepts are constituents of thought, and make a stable contribution to the content of the thoughts in which they figure. When a subject thinks about an object in the world and its properties, he does so by instantiating concepts, which are semantically related to objects and properties. Concepts are seen as psychologically real, and roughly analogous to

[1] Ch. 2, ss. 3.5, 4.2 & 4.3; ch. 3, sec. (4).

linguistic words. The thoughts in which they figure can be true or false, and the semantic relation of concepts to things in the world accounts for the fact that the conditions under which a thought is true can differ from the conditions in which a thinker would entertain it. People may sometimes judge falsely. Thus, a subject's categorisation judgements may be false. The things he categorises as falling under a concept may not in fact satisfy it. Thus, the extension of a concept is not simply determined by a complete characterisation of the way a subject would apply it.[2] That is not to say that features of categorical practice have no role to play in determining extension. But, if concepts have a reference-determining role, then reference cannot follow simply from the categorical practice. And experimental psychologists need this distinction, because it forms the basis of their measure of accuracy. If a subject has made 87% correct and 13% incorrect categorisations, then there must be some fact of the matter about the correctness and incorrectness of his application of the concept which outstrips the way he actually applies that concept in categorisations.[3]

It is easy to confuse subjects' use of a concept to make category judgements with the referential properties of that concept (since we can also say that, by referring, the concept divides the world up into categories). To be clear, I will reserve 'categorisation' for subjects' actual applications of a concept, to distinguish it from a concept's referential role.

Finally, I should dispose of a preliminary objection to the discussion of prototypes in this chapter. Prototype theories are theories of concepts, but I have argued that clusters in connectionist networks are not concepts and do not have conceptual structure. How, then, can connectionist clusters be relevant to prototypicality? The answer follows from the explanation, in subsection 2.1 of chapter 3, of how connectionist clusters might

---

2 Usually, this would be expressed as the claim that contents are verification-transcendent, as indeed I believe they are. However, even a verificationist has the resources to make a distinction between actual uses and correct uses (Dummett 1976). The trap I am adverting to is the temptation to a very strong verificationism, sometimes found amongst experimental psychologists, which assumes that the reference of a concept consists just of all and only those things that a person would categorise under it.

3 A theorist who wanted to hold on to prototypes as determinative of reference might argue that a subject's rapid categorisation judgements do not determine the extension of his concept, but that the extension is fixed by the underlying prototype which drives those categorisations. Alternatively, the theorist could argue that categorisation judgements do determine a concept's extension, and allow that everyone's concepts have slightly different extensions. If so, 'correct' and 'incorrect' would mark the extent to which a subject's idiosyncratic extension matches some general trend.

act as inputs to a system of conceptual representation, or might form part of such a system. Recall that, according to my theory, clusters have complete contents[4] – they make claims about the properties of the presented sample: P(s). The only aspect of that claim which varies, as different clusters are activated in a given state space, is the property represented. So the state space keeps track of those properties. That would allow it to act as the basis for a categorical judgement, the most basic of which is simply: *that is P*. Thus, some of the typicality effects found in human *conceptual* systems could be due to the operation of connectionist components.[5]

The remainder of the chapter is structured as follows. Section (2) sets out a basic version of one prototype theory of concepts, as a framework in which to explain the psychological evidence for prototype effects, which I do in section (3). There is a convincing body of experimental evidence that our conceptual abilities do indeed display typicality effects (subsection 3.1). I also mention the evidence for the existence of a basic level in the hierarchy of categories (subsection 3.2). Section (4) explains three specific theories of what prototypes must be, in order that they fulfil the task of underpinning abilities to categorise. There are three broad varieties of theory, with prototypes as: feature lists, exemplars, and stored perceptual representations, respectively. The first holds that prototypes are lists of features which are statistically reliable properties of the category in question. The second has a prototype as a group of exemplars, which might include representations of individuals. This type of approach is closest to the Wittgensteinian suggestion that the extensions of concepts are groups of instances formed into an equivalence class in virtue of pairwise family resemblance relationships. The third broad type of approach has a concept as one or more stored non-conceptual perceptual representations.

Sections (5) and (6) canvass objections to the claim, central to most prototype theories of concepts, that prototypes determine content. Section (5) discusses how prototypes combine in complex concepts. Various positive claims are made in the literature about prototype combination, but worries about conceptual combination motivate an objection to prototype theories of concepts. In section (6) I consider four further objections to prototypes as content-determining: circularity / regress (6.1), 'well-

---

4 Ch. 2, ss. 3.5 explains why they should have complete contents.

5 For further details, see ch. 3, ss. 2.1.

defined' concepts (6.2), ignorance and error (6.3), and concept stability (6.4). The conclusion to be drawn from the objections in sections (5) and (6) is that prototypes do not determine content. Section (7) shows that prototypes need not be psychologically real at all, in some cases. Connectionist models show how typicality effects can arise in a system that does not store prototypes. Furthermore, I explain why such typicality effects occur, relying upon the clustering approach from chapter 2.

(2) A BASIC PROTOTYPE THEORY

As a starting point I will set out a theory of concepts as prototypes based roughly on Rosch (1978).[6] This will provide a framework for considering more refined and alternative approaches, as well as for assessing the attractions and drawbacks of these theories. The basic idea is that concepts are lists of features each of which is a statistically reliable attribute of the referent. This can be seen as a variant on the classical definitional view of concepts. Concepts remain structured lists of features, but the requirement that they apply to everything in the extension of the concept is relaxed. So the theory is that concepts are structured mental representations that encode the features that objects in their extensions tend to possess. These features are elucidated experimentally in the first instance by asking subjects to list the properties they associate with a given category. Those features are themselves concepts, so there is an obvious worry about regress or circularity. Typically, theorists assume that feature associations bottom out somewhere (subsection 6.1 below considers whether this is plausible).

So the prototype for the concept is a list of features. It is applied to objects by judging the similarity between the representation produced by an object as it is experienced, and the prototype. Any object producing a representation similar enough to the prototype falls in the extension of the concept. How is similarity determined? One common means is to use the Tversky (1977) contrast principle. The similarity between a

[6] Rosch (1978) explicitly disavows taking her description of prototypicality effects in experimental settings as entailing any particular theory of how categories are represented. However, in Rosch & Mervis (1975) and Rosch (1977) she suggests that one reasonable theory suggested by her findings is that a category is represented by the prototype that is most representative of the items in the category and least representative of the items outside the category. The discussion herein reformulates that claim in terms of how the extension of a concept represented by a prototype must be determined in order for it to hold true.

representation of an instance and the concept C (also a representation) is taken to be a weighted sum of the common features less the distinctive features:

Sim (I,C) = af(I∩C) - bf(I-C) - cf(C-I)

Where:-

I is a mental representation of an encountered instance, consisting of a feature list

C is the concept (a representation), also a feature list

Sim (I,C) is the similarity between I and C (a numerical measure)

I∩C are the features common to I and C

I-C are the features of I not found in C

C-I are the features of C not found in I

f is some appropriate function. Most simply, it will just count the number of features which are shared and distinctive

a, b, c are (positive) weights of the relative importance of shared and distinctive features

The relevant function f must be assumed and the parameters, including the importance to be assigned to each of the attributes, must be determined from empirical data. An example illustrates how this is supposed to work. Suppose the prototype for APPLE is something like: ROUND, GREEN OR RED, GROWS ON TREES, TASTY, etc. Then a yellow Golden Delicious counts as an apple because it has many features in common with the prototype and only one distinctive feature, its colour. By contrast a red ball does not count as an apple, since its attributes are more distinctive than similar to the important attributes of the apple prototype. The idea is that a similarity space around the prototype determines the concept's extension: anything which reaches a high enough similarity score is in, everything else is out. On a common reading, membership of the category is taken to be graded, depending upon the similarity scores. Many instances will be clearly highly similar (in) or highly dissimilar (out), but other instances will lie on the borderline, and so are intermediate members of the category.

The theory also explains how concepts enter into inferences: inferences are based on the structure of the prototype. Pursuing the foregoing example, on the basis of the prototype for APPLE an thinker can infer from *that is an apple* to *that is tasty*. These inferences will be non-demonstrative, but that is a virtue in the light of the difficulties classical theories of concepts face in isolating analytic connections as the basis of

demonstrative inferences. Furthermore, the structure of the prototype informs the inductive strength of these inferences. There is also a natural model of concept acquisition which fits with the prototype story: to learn a new concept is to acquire its prototype, by means of collecting together statistically reliable features. As we have seen, prototype representations allow for a graded notion of category membership, based on degree of similarity to the prototype. They are also compatible with strict category membership, with some cut-off point. However, their greatest virtue is not that they allow for graded category membership, but rather that they have a structure within which instances can be more or less typical. It may be that even un-typical instances should be considered as full members of the category. But it is the typicality space within the structure of a concept which prototype theories aim to capture. And as we will see in the next section, there is strong empirical evidence that many concepts have an associated typicality space.

There is a further claim about conceptual structure in Rosch (1978) which is less frequently considered. That is the idea that there is a hierarchy of concepts, and that within this hierarchy a level of basic objects can be discerned. The basic level is the most inclusive level of categorisation at which objects in the category have a relatively large number of attributes in common. Think of this roughly as the most inclusive level at which a genuinely representative member of the category can be found. For example, classifications of artefacts form a hierarchy. One category is furniture, which divides into chair, table, bed, etc. Each of these divides in turn, e.g. chair into dining chairs, easy chairs, etc. The claim that there are basic level categories is the claim that, within this hierarchy, there is a privileged level. In this case, the basic level categories are chair, table, bed, etc. That is the most inclusive level at which genuinely representative members of the category can be found. Furniture is superordinate and types of chairs are subordinate to this basic level.

Rosch formulates the measure 'cue validity' as a means of discerning the basic level categories (Rosch, Mervis et al 1976). The cue validity of a particular attribute for a given category is the conditional probability that an object falls within the category given that it has the attribute. It is a measure of how good a given attribute is as a predictor of that category. Formally, for given category C the attribute x_i has cue validity = $P(C|x_i)$. This obviously increases as x_i is more reliably a property of objects in category C. But it also decreases as x_i is also associated with other categories, and thus fails to be diagnostic

of C. The total cue validity for a category is the sum of the cue validities of all the attributes in its prototype:

$$\text{Total cue validity} = \sum P(C|x_i)$$

Rosch's claim is that there is a basic level in which the categories have a higher cue validity than both superordinate and subordinate terms: the former because higher level categories do not reliably possess a fixed set of attributes; and the latter because lower level categories have few diagnostic properties. The same property can also be formalised in terms of Tversky's measure of category resemblance discussed above.

(3) THE EMPIRICAL EVIDENCE

3.1 Evidence for typicality effects

Four main types of evidence support the claim that concepts have prototype structure:-

(i) graded judgements of typicality;
(ii) typicality assessed from spontaneously listed features;
(iii) speed of quick categorisation judgements;
(iv) categorisation errors.

The first category of evidence comes from asking subjects how typical a particular instance is as a member of a given category. This is usually done just using words, e.g., 'how typical a bird is a chicken / an ostrich?' It has also been tested using pictures. The striking result is that subjects agree about how clear a case of a given category each instance is (Rosch 1974; 1975, p. 197). This is usually demonstrated by asking subjects to rank instances for typicality. They are found to agree on their rankings. This agreement remains even if the subjects disagree about where the boundary for the category should be drawn.

These graded judgements of typicality also show up in linguistic substitutability and in what Lakoff (1972) calls 'hedges': qualifications like 'almost'. Rosch (1977) shows that the appropriateness of substituting a subcategory in a sentence aligns with the typicality ratings elicited by asking subjects to grade typicality. Thus, 'bird' can be more readily substituted by 'sparrow' than by 'penguin' in a sample of linguistic contexts. Rosch also suggests that hedges like 'almost', 'virtually' and 'technically' are more readily

applied when an instance is an un-typical member of a category. These linguistic data seem to be just an application of the result elicited by asking people directly — namely, that some members of a category are more typical than others.

It is tempting to conclude from this agreed graded typicality that membership of a category is graded. Indeed, that was an early interpretation of these experimental results. However, that interpretation remains controversial. As we will see below (subsection 6.2), even well-defined categories like *odd number* generate these typicality effects. Three is agreed to be a better example of an odd number than is ninety-one; yet subjects agree that whether a number is odd is an all or nothing affair. So the existence of graded typicality as part of the structure of a concept does not imply that membership of that category is graded.

The second way of assessing typicality is by asking subjects to list the attributes of various categories and instances. For example, subjects are asked to list the attributes of the category *bird* and of various category members, like *robin, sparrow, hawk, ostrich*. Each category is assumed to have, more or less reliably, the features listed, with the features listed most commonly by the subjects as the most important or reliable. These lists are found to predict typicality judgements: the more attributes that a category member has in common with other members of the category, and the fewer attributes it has in common with contrasting categories (using Rosch's cue validity and Tversky's contrast principle), the more typical of that category will it be rated (Rosch and Mervis 1975). So for example, *robin* but not *ostrich* comes out as typical of *birds*. Furthermore, the features which the different members of a category have in common are the same features as will be found in the list given for that category. In our example, the features listed for *bird* will correspond to those features common across *robin, sparrow, hawk, ostrich*, etc. This is then taken to explain the typicality judgements: ROBIN has more of the features of BIRD than does OSTRICH, and so *robins* are rated more typical of the category. The typical members not only have the most commonly listed features, but where those features have a metric, correspond to averages along the dimensions of those features. So if category members vary in size, say, the typical members will be of a size that is the statistical mean or mode of the members (Reed 1972, Rosch, Simpson et al 1976).

A second group of experiments within the paradigm of spontaneous categorisation relies on subjects producing lists of category members. Subjects are asked to list members of some category. Those that were most frequently listed were taken to be typical

members of the category. These results agreed with those from the other measures of typicality discussed above (Rosch 1975, Rosch, Simpson et al 1976).

The third type of task on which typicality is tested is speed of categorisation. One paradigm is to ask subjects to respond with "true" or "false" to statements of the form: X is a member of Y. Reaction times are found to correlate with typicality (Rosch, Simpson et al 1976). The intuitive idea is that, the more similar the representation of an instance is to the prototype for a category, the easier and thus quicker it is to classify as a member of that category. However, care must be taken in translating this intuition into a processing model. For example, in the basic model discussed in the last section, an instance is graded for typicality based on Tversky's contrast principle. Recall that this involves taking a weighted sum of the common characteristics less the distinctive characteristics. To arrive at the correct sum thus requires all the characteristics to be checked. There is no short cut when the represented instance has a lot of features of the prototype; they may still have many distinctive features. So there is no quick categorisation whereby once the two have enough in common the instance is definitely 'in'. There is no obvious reason why calculating a Tversky-type function should be any quicker for typical cases than for non-typical cases or for typical non-cases. Rosch's cue validity has more potential to explain the reaction time phenomenon. If her prototypes are taken to consist of a list of features for a category together with their cue validities, then a natural model of category processing does suggest variations in reaction time. If a represented instance has many of the features which are not only common to the category, but also good predictors of that category, then these features will quickly reveal that the instance is likely to be in the category.

Thus, the reaction time experiments show that care must be taken in interpreting results. Not every experimental paradigm whose results correlate with typicality tends to demonstrate that concepts have a prototype structure. It was only on one particular model of category processing that the reaction time results support a prototype structure. Another equally compelling explanation for the results is that they reflect some rough and ready epistemic rules for judging category membership, based on experience with the category or knowledge about that category and related ones. In which case, a prototype is not part of the structure of the concept (not part of an account of what makes it the case that some instances are members of it and others are not), but graded speed of categorisation based on typicality is still predicted. Rosch, Simpson et al (1976) attempts to block this conclusion by using artificial categories and controlling for frequency of

experience with the items used. This does suggest that it is not some brute associationism based on mere frequency of experience with individuals which underlies speed of categorisation. However, it still does not rule out the reaction time results arising from epistemology of the category – how subjects judge what is in and out – rather than from the structure of the concept itself.

The fourth experimental paradigm in which typicality effects are demonstrated studies categorisation errors. The experiments are similar to those just described, but accuracy rather than speed is the variable under investigation. The results show that categorisation errors are inversely correlated with typicality. As with the previous experiment, care must be taken in assessing these results. On the Tversky model of similarity, all features have to be compared before a similarity measure can be calculated, so there is no reason to suppose that more errors should be made with classifying non-typical than with typical instances. However, a simple accumulator model of categorisation would explain these results for category membership (but not non-membership). In this model, features are compared between the prototype for a category and a represented instance to be assessed for category membership (along the lines of Smith & Medin 1981, and Smith 1995). As features are compared, common features add a weighted sum to an accumulator; once a high enough score is accumulated, the instance is judged as falling within the category. This would explain why typical instances are categorised more accurately: their representations are very like the prototype, so few comparisons are needed in order for the accumulator to reach a level where they are judged to be in the category. Notice again that care is needed here. If the accumulator adds for common features and subtracts for distinctive features, then there is no reason to suppose that the score, once having crossed the relevant threshold, will not subsequently fall below it again due to distinctive features. So on such a model, all the features must be compared if typicality effects are to be generated. Perhaps such features are compared in parallel. Even so, there is no obvious reason why accuracy should be affected by similarity to the prototype. Furthermore, other experiments suggest that accuracy and speed are both high for clear non-members. On the accumulator account, the conclusion that something is a *non-member* would seem to require comparison of all the features.

This is not to claim that the results concerning speed and accuracy of categorisation (in conditions of time-pressure) cannot be explained by an appropriate model of the structure of a concept. However, many of the objections are avoided by taking the approach suggested above. That is to suppose that these results arise from the

mechanisms by which instances are judged to belong to a category or not, not from that which determines category membership. It may be epistemically easier and quicker to judge typical instances as in the category; thus producing fewer errors, especially when working under time pressure. This does not obviate the need to explain how in processing the category judgement, speed and accuracy are correlated with typicality. However, it removes the constraint that whatever explains these patterns must be part of the reference-determining structure of the concept. So it allows experience, real world knowledge about related categories, and about their features, to play a role in the epistemic task. It may be that this is what is reflected in the typicality effects elicited.

Support for this conclusion is found in Landau (1982), where he demonstrated that young children switch criteria of categorisation when the task is changed from identification to justifying their categorisation. This suggests that an operational difference can be drawn between the means of identifying which category various instances belong to, and what makes it the case that certain instances belong to a given category. The experimental paradigms described above are not able to discern whether the typicality effects they elicit are features of the former or the latter. So, whilst these robust experimental effects require an explanation, that explanation may ultimately be found in epistemology of identification rather than the metaphysics of the content of concepts.

3.2 Evidence for basic level categories

The first claim in common to prototype theories of concepts is that concepts have a prototype structure, in a way which explains the experimental results discussed above. What about the second claim: that concepts form a hierarchy, in which can be ascertained a level of basic concepts which have some kind of privileged status with respect to their subordinates and superordinates? Rosch (1978) marshals some evidence for this conclusion. First she investigated whether a basic level concept could indeed be found using the criteria suggested by Berlin (1978). Subjects were asked to list attributes for each of several categories in a hierarchy (eg, furniture; chair, table, …; kitchen chair, …). She found that at the highest level of categorisation (e.g., furniture) subjects listed relatively few attributes, at the next level down (e.g., chair) listed many more attributes, but few additional attributes were elicited by moving further down the hierarchy (eg, kitchen chair). So it is apparently possible, within a hierarchy of categories, to identify

the most inclusive level at which objects have many attributes in common (in this case, chair), i.e., the basic level.

Rosch repeated this test on motor movements (albeit only on subjects' descriptions of motor movements). She asked subjects to describe in as much detail as possible how they would interact with objects in various categories. These responses were classified into movement classes. It could then be asked of a category: does it have many motor movements in common? The most general categories where the answer to this question was "yes" agreed with the basic level categories elicited in the previous experiment. For example, there are relatively few motor programmes that apply to all pieces of furniture, but relatively many that apply to all chairs. Interacting with kitchen chairs and dining room chairs uses essentially the same motor movements as for chairs in general. Obviously, these results can be questioned. They are based on subjects' descriptions, in a relatively limited set of taxonomies, and depend heavily upon how it is decided to classify the motor programmes. However, they are suggestive of the existence of a basic level.

Rosch carried out some further, if questionable, studies to test her hypothesis of the existence of basic level categories. This made use of two-dimensional outlines of instances of the objects in the category (e.g., of cars, or of vehicles in general). She does not describe how these outlines were chosen so as to be representative of the category. She analysed the ratio of overlapping to non-overlapping areas in the shapes. She found that outlines in superordinate categories (e.g., vehicle) overlapped relatively little. There was then a big jump to objects in the hypothesised basic level categories (e.g., car), which overlapped relatively well, and only marginally less than the outlines in the subordinate categories (e.g., sports car). Using these same sets of shapes Rosch constructed average shapes for the categories (again, it is not clear what criteria were used to produce the 'averaged' outlines). Unsurprisingly, subjects were unable to identify the composite objects at the superordinate level (e.g., vehicle). The most general level at which they could identify the object depicted was at the hypothesised basic level (e.g., car).

Rosch (1978) argues that there is further evidence supporting the existence of basic level categories in work on imagery, perception, category development and language acquisition. She claims that the basic level is the most abstract category for which an image of a typical member can be representative of the class as a whole; and that objects are first perceived as members of their basic level category and only later identified under their super- and subordinate level categories. She gives no details of these experiments, so it is hard to assess them. Furthermore, in children's development of categories, Rosch

claims that basic level categories are the first named, and the first basis for sorting objects prior to naming. Finally, using American Sign Language as a model for language acquisition, Rosch, Mervis et al (1976) claims that basic level categories are more likely to be coded by a single sign than are super- or subordinate categories.

All these experimental results are suggestive of the existence of a privileged level of categorisation. However, even if they are taken at face value, they do not establish that a category hierarchy is part of the structure of those concepts. Although some of the experiments attempt to control for frequency, they do not rule out the effects arising from some combination of familiarity, knowledge and usefulness of the categories in question. The epistemology of these categories may need to show how, in placing an instance in a category, basic level categories have some kind of priority. It is not clear at all that such effects form part of the correct account of the structure of a given concept: of what determines its content. Even more than typicality effects, the privileged status of basic level categories appears to be a phenomenon that overlays the conceptual structure.

Nevertheless, if real, it is an important psychological phenomenon. Rosch gives some reasons above why such a privileged level may exist. A fully worked-out theory of concepts would explain in what this hierarchical structure inheres, by what mechanism it arises, and would give a developmental or evolutionary account of the preferential status of concepts of the kinds of things which form basic level categories. In section (7) below, I suggest how connectionist models may account for something of this hierarchical structure. That is unlikely to be the whole story. However, provided the basic level phenomenon is not thought to play a role in determining the content of concepts, its existence does not rule out the applicability of the connectionist model. Rather, it adds another explanatory task if clusters in connectionist state space are seen to implement part of a conceptual system.

(4) SOME VARIETIES OF PROTOTYPE THEORY

We have seen in subsection 3.1 above that there is strong experimental evidence for the existence of typicality effects in categorisation tasks, and in subsection 3.2 that there is some evidence for a privileged basic level of concepts. That much is relatively uncontentious. So-called prototype theories of concepts attempt to explain these data as stemming from the fact that concepts have a prototype structure. In fact, there is no one settled prototype theory of concepts. In this section, I will set out three broad varieties of prototype theory.

The first is the type of theory already elucidated, using feature lists. A category is represented by a prototype. The prototype consists of a list of features which are reliably associated with that category. The concept consists of the prototype together with some similarity metric, which determines when represented instances are similar enough to the stored prototype to be within the extension of the concept, and when outside it. So the prototype is an abstract summary of the features related to the category, and need not relate to any individual instance.

Exemplar-based theories such as that of Smith & Medin (1981) allow all or part of the concept to consist in representations of particular instances of items falling under the category. For example, a representation of your childhood pet might be your prototype of the concept DOG. The difference between exemplar-based views and feature lists is a matter of degree. The exemplars which form the structure of a concept must themselves be represented somehow. In Smith and Medin's (1981) theory these exemplars are ultimately represented in terms of lists of features. The difference from the pure feature list view is that these features characterise not the category as a whole, but some subcategory or particular instance. In the most basic case, a single exemplar would represent the whole category. This might be a subcategory or an instance. So, for example, the category *bird* might be represented by the exemplar ROBIN which is picked out by a feature list (ANIMATE, FEATHERED, RED-BREASTED, etc.). Or the category might be represented by a particular instance, your pet bird "Fluffy", represented by a different set of features (ANIMATE, FEATHERED, YELLOW, CAGED, etc.). However, in the general case more than one exemplar will stand for a given category: so *bird* might be represented by a mixture of the subcategories ROBIN, BLUEJAY and SPARROW, together with the instance "Fluffy". And there is no limit on how deep the structure might go: subcategories might in turn be represented by a mixture of particular instances and sub-subcategories.

So the distinguishing feature of exemplar-based theories is that they allow representations of particular instances to form part of the structure of a concept. Feature lists are in general more abstract representations of the category than are either the subcategories or particular instances allowed on the exemplar view into the structure of a concept. In addition, exemplar views allow more than one subcategory / instance to form part of the concept. In processing a particular categorisation not all of these exemplars may be used. In a feature list representation, in general all the features in the list are employed in making a categorisation. This makes exemplar views needlessly disjunctive: the multiple exemplars and subcategories contain overlapping information. In the

proximity model, which is the most extreme case, each concept is represented by all instances that have ever been encountered Reed (1972). This would seem to place excessive demands on memory and processing capacity to be psychologically realistic.

The proximity model also raises starkly a question which must be answered by any exemplar-based view: in virtue of what are the represented instances and subcategories drawn together as forming a single concept? If they are collected together in virtue of something they have in common, then surely those common features (whether they are definitive or merely statistical) are what the actual structure of the concept consists in. There must be some prior way in which the exemplars are collected if they are to determine the content of the concept. One suggestion is that the examples are taught: we store a set of exemplars as we are taught which are good examples of a concept. Another approach allows the best examples to change: an initial set of exemplars determines, of various instances, whether they fall under the concept or not. But then the exemplars are re-calibrated so as better to represent the range of instances actually encountered. That would allow the exemplars to change without necessarily altering the content of the concept.

Whether the exemplars are taught, or chosen and recalibrated, the idea of this sort of view is that a set of *best examples* forms the concept. There are at least two proposals of how categorisation judgements are then processed. One idea is that the representation of a new instance must match sufficiently (in its relevant features) at least one of the exemplars (subcategories or instances). A second proposal is that an instance operates to 'retrieve' exemplars based on general similarity constraints, where the exemplars are chosen from all those found in any concept. The instance is then categorised as falling under a concept C if it retrieves n (for some n) exemplars of C before it retrieves n exemplars of any other concept. Clearly, both of these proposals would need expanding to turn into testable empirical hypotheses about categorical processing; but they serve to illustrate the idea.

Baldly stated, the exemplar view also lacks an explanation of how the similarity space around a given exemplar or set of exemplars is determined. Recall that feature list prototype theories had to make a similarity metric part of the structure of the concept: a way that represented instances are judged as similar or dissimilar to a prototype, and features are assigned more and less importance for a given concept. The exemplar view must do something similar: representations of exemplars alone will not fix what counts as

a sufficient match to a given exemplar; which is the starting point for determining whether the represented instances match enough of the exemplars of a concept.

Medin and Schaffer's (1978) context model has a partial answer to this question. They assume that a thinker must be taught or otherwise learn that some represented instances fall under the concept. The rough idea is that this set of instances will determine what are the most salient features for that category: the ones which most instances have in common. Then exemplars are stored on the basis only of those features, with the more typical examples being given more weight. It is called the context model, since the model envisages context of categorisation as a further variable which influences the chances of retrieving individual exemplars for comparison, thus allowing context to affect categorisation decisions. The model is a puzzling combination of weighted feature and exemplar views, and seems empirically implausible as a model of concept acquisition. However, it does point up the fact that the end state should include not only exemplars, but means of telling what are the important features for judging similarity between exemplars and represented instances.

One nice advantage of Smith and Medin's exemplar view is that it explains some interesting similarity relations. For example, CHICKEN is judged more similar to ANIMAL than to BIRD, whereas ROBIN is judged more similar to BIRD than to ANIMAL. According to the exemplar view, this is an immediate consequence of the fact that CHICKEN is an exemplar of the category *animal* and ROBIN is an exemplar of the category *bird,* but not the converse in either case.[7] Similarly, opponents of the idea that prototypes form part of the reference-determining structure of concepts have space to argue that these similarity effects arise either from some other feature of the structure of concepts, or that they reflect the information associated with these concepts.

The third broad type of prototype model is much more parsimonious. It is an exemplar-type view, but with exemplars as non-conceptual representations. The idea is that a concept consists of one or more stored (roughly) perceptual representations, rather than consisting of semantic information. Any represented instance sufficiently similar to the image/sound/texture etc. counts as falling under the concept. As before, the exemplar must be supplemented by some similarity space if it is to represent a category. However, that feature of the concept may be an automatic consequence of the perceptual representation. So, if only shape and sound are stored, then only these will be the basis of

[7] Other prototype theories have less elegant means to account for this finding.

comparison. There must still be a metric of comparison (for example, comparison by size) but that might be achieved at a relatively perceptual level, without needing complex semantic machinery. This proposal has the merit of illustrating how concepts might be formed by employing non-conceptual resources, which is a singular failing of many prototype theories. However, it remains only a sketch.

Wittgenstein's observation that some categories are formed out of family resemblance relationships inspired some of the early work on prototypes. In particular, his rejection of a classical definitional theory of concepts left theorists looking for an alternative. His observations seem most naturally compatible with an exemplar view of one of the two varieties just mentioned. I suspect he would be antithetical to lists of features circumscribing the content of a concept, even if they are statistically reliable rather than defining features. This is an important observation, since 'family resemblance' is often taken to mean having sufficiently many of some weighted set of features. In particular, that is how Kripke reads this type of proposal in attacking description-based theories of content in lecture II of *Naming and Necessity* (Kripke 1972). However, part of Wittgenstein's insight is that there need be no features in common between all or even most members of a category. Consider an exemplar model of stored images. Suppose that to count as falling under the concept a represented instance has only to be sufficiently similar to one of the stored images (exemplars), along any of the dimensions represented in that image. So an instance will count as 'in' if it is sufficiently like image A in shape, or image B in colour, or image C in size, etc. Then there need be absolutely nothing in common in the represented properties of the instances which fall under the concept. What then unifies them as falling under the same category? Well it would be sufficient if there were pairwise resemblance relations between the different exemplars. Eg, image A is similar to image B in colour, image B to image C in size, etc. In that way the set of exemplars determines the boundaries for a concept which could not be fixed in terms of common features. Nor need this imply that the content of the concept must be arbitrary: there may be something in common in nature between the referents of the exemplars which represent the concept, even if nothing is common in the stored representations. To form a coherent category all that is required is that the pairwise resemblance relations between the exemplars form a rough equivalence class. That will allow the concept to be distinguished from its negation. And there is no reason to suppose that such equivalence classes must be based on using the same criteria of comparison between all the members. So 'family resemblance' should not be taken as synonymous with weighted feature lists.

This section has only given an indication of the range of different views which fall under the rubric 'prototype theories of concepts'. Next I will consider some objections, many of which apply to all or most of the types of theories discussed. However, I do not claim that any theory is susceptible to all the objections I will raise, nor that any one of the objections applies to all the theories, thus succeeding in knocking out any prospect of a successful prototype theory of concepts. Rather, the objections push prototype theories in a certain direction, which I will claim makes it more plausible that typicality effects are to be accounted for otherwise than by virtue of the reference-determining structure of concepts.

(5) COMBINING PROTOTYPES

Concepts combine to form complex concepts. According to definitional theories, the content of a complex concept is determined in the same way as the content of its constituents: by its associated definitions. That story does not work so well if prototype structure determines content. Accordingly, it often thought that prototype theories have a special problem with conceptual combination. For example, it is Fodor's principal objection to prototype theories in his polemical Fodor (1998). However, the problem is better read as a request for clarification: how do prototype theories account for conceptual combination? The purpose of this section is first to examine some of the proposals for answering that question, and second to consider how questions of conceptual combination put prototype theories under pressure.

It is almost universally agreed that concepts can combine to form complex concepts. For example, RED CUBE is formed out of RED and CUBE. Furthermore, RED CUBE seems to inherit its content from the content of its component concepts RED and CUBE (whether or not those components are themselves complex or atomic). Furthermore, it is widely thought that genuinely complex concepts get their contents compositionally from their components. That is to say, RED makes the same contribution to the content (application conditions, say) of all the complex concepts in which it figures. Thus, RED plays the same role in RED CUBE and RED SPHERE. Contrast the area of Moscow called 'Red Square'. The concept RED does not play any role in the content of the corresponding concept REDSQUARE.[8] Indeed, on many views the concept REDSQUARE is likely to be atomic.

[8] At least, no direct role. On some prototype views REDSQUARE may be a highly complex concept, and RED may occur in the prototype, perhaps as a feature of the old Soviet flag.

So the question for prototype theorists is: what do you say about (genuinely) complex concepts? Do they have prototypes? If so, how are their prototypes formed? If not, how is their content determined?

Two sources of data inform this debate. The first is that some complex concepts *do* appear to have prototypes, but that these are not simple conjunctions of the prototypes of their constituents. To use Fodor's favourite example, suppose the prototypical *pet* has roughly the features of a dog (medium sized, trainable, furry, pet-able, eats anything, etc.) and the prototypical *fish* is roughly like a salmon (long, slimy, scaly, water-living, etc.), then a simple combination of these prototypes does not produce a suitable prototype to associate with the concept PET FISH (*not* long, medium sized, slimy, trainable, scaly, pet-able, water-living, omnivorous, …), or for anything at all. A prototype for PET FISH should have roughly the properties of a goldfish. Where does this prototype come from?

The other source of data is that there are innumerable complex concepts for which there appear to be no prototypes. Laurence & Margolis (1999) list several plausible examples:-

(i) Uninstantiated concepts: U.S. Monarch, 31st century invention.

(ii) Negations; NOT A WOLF (or other Boolean constructions, like conditionals).

(iii) Heterogeneous categories: FROG OR LAMP, NEW SPECIES.

(iv) Wide categories: A CONSEQUENCE OF AN ONGOING PHYSICAL PROCESS SOMEWHERE IN THE UNIVERSE, OBJECTS WHOSE MASS EXCEEDS 1KG.

(v) 'Intricate' concepts like BELIEF.

These are all cases where we seem to possess a concept without having a corresponding prototype.

Prototype theorists are under no obligation to claim that all concepts have a prototype structure. (Indeed, the prototypes must bottom out at some point in primitive components whose content is determined in some other way.) So, if experimental investigation shows that concepts like those listed by Laurence & Margolis do not have prototypes, then prototype theorists can deny that those concepts have their content determined by a prototype structure. Why should that be? Well, for many of the examples given, it is plausible that the concept's content is determined from the composition of some constituents. So the prototype theorist can deny that complex concepts have their contents in virtue of prototypes.

This also allows the prototype theorist to deal with the first class of case, where complex concepts seem to have the wrong prototypes. She can simply stick to the view

that complex concepts do not get their content from their prototypes. How do they get their contents, then? The standard way: as a compositional consequence of the contents of their constituents. This leaves the prototype theorist free to argue that some of those constituents do have content-determining prototypes. Content need not be determined in the same way for lexical concepts as it is for concepts that are formed from combinations of lexical concepts.[9] According to this line of response, if a subject needs to judge whether some object O falls under the concept PET FISH the prototypical *pet fish* (or the absence of one: see NOT A WOLF) is irrelevant. Rather, using the prototype for PET he judges whether O is a PET, similarly for FISH, and if it passes both tests, it satisfies PET FISH.

That is a perfectly adequate line of response. However, this is one area where prototype theories lose out to the rival classical definitional theories. To see why, notice that according to definitional theories lexically complex concepts refer in exactly the same way as lexically basic items (in fact, both are complex, according to definitional theories). In each case, they refer in virtue of their definitions. PET FISH = PET & FISH = DOMESTICATED & … & SLIMY & …, and so on, until decomposed into the primitive constituents, whether these be sense data or primitive concepts of things in the real world. At first sight, the fact that prototype theory needs to combine the contents themselves, and not the means of determining content, seems slightly unattractive when compared to the elegance of the classical theory. But on careful consideration, there is nothing strange about arriving at the content of a complex concept via a compositional principle — breaking the concept into its conceptually primitive components and judging whether each one applies by means of their prototypes.

Interestingly, Fodor (1998) thinks that the major objection to prototype theories is that prototypes don't compose. This is puzzling, since Fodor's theory is that the content of a concept is fixed by its informational relations (that informational relation upon which the others asymmetrically depend gives the content), and these don't compose either.[10] He argues that most lexical concepts have their content in virtue of asymmetric dependence relations. But he agrees that asymmetric dependence relations don't compose. They don't need to because, according to Fodor, genuinely complex concepts have their contents definitionally, by combining the contents of their constituents. The prototype theorist can say the very same thing. I can't find any explanation from Fodor

[9] On most theories it would not be; cf. the demand that it should in Werning (2003).

[10] Laurence & Margolis (1999) also point this out.

why the approach he advocates for his asymmetric dependence relations is not equally good for prototypes. It seems to me clear that his own approach provides the prototype theorist with a perfectly coherent account of compositionality.

However, there is still a worry to be found in the data about complex concepts; but it is certainly not the kind of knock-down blow which Fodor supposes. The remaining worry lies in the fact that some complex concepts *do* seem to have prototypes.[11] Intuitively a goldfish is a prototypical *pet fish*. Complex concepts do not have their content fixed by prototypes, but some such concepts still have associated prototypes. That substantially weakens the inference from the existence of typicality effects to prototypes forming part of the reference-determining conceptual structure – it is admitted that some complex concepts have prototype structure, but that it is not reference-determining. So the typicality effects must be due to some other feature of the conceptual set up: epistemic tests, real world knowledge, familiarity, etc. Which makes it more plausible that the very same story explains the typicality effects found with lexical concepts, namely, that they are due to something associated with a concept that is not part of its reference-determining structure.

(6) OBJECTIONS TO PROTOTYPES AS CONTENT-DETERMINING

6.1 Circularity / Regress

In the last section, we saw that complex concepts gave us a model where prototypes are psychologically real, but not content-determining. That weakens the claim that prototypes determine the content of any type of concept. In this subsection, I will set out four further objections to the idea that prototypes determine content. Taken together, they present a strong case that, despite the robust reality of typicality effects, that which accounts for them is not determinative of content.

Both the feature-list and the standard variety of exemplar views of prototypes require that further concepts can form constituents of a given prototype. On its face, the theories thus face the same objection of circularity or regress as do classical definitional theories: if one prototype is constituted by other concepts, which in turn have their contents fixed by prototypes containing further concepts, how does the whole system

[11] The experiments reported in Smith, Osherson et al (1988) do suggest that some complex concepts, like RED APPLE, have associated stereotypes.

achieve any anchor in the world? The type of response is the same as that made by classical theorists – concepts form a chain of increasing primitiveness which eventually stops somewhere. At some stage there are primitive concepts which are brutely available to the thinker, and the thinker is so set up as to employ these concepts with a certain reliability. According to empiricists, these concepts must be sensory, for example representations of how various aspects of the visual field appear. Current theorists broaden the perspective and allow that the primitive base include world-involving concepts. But in either case, the chain of concepts found in prototypes has to end with plausibly primitive concepts: concepts whose identification procedure does not involve the application of any further concepts. It may be that prototype theorists can show how prototypes 'bottom out'. Some proposed prototypes seem to be moving in the right direction, eg, APPLE whose prototype is given in term of colour, shape and texture attributes (Laurence & Margolis 1999, p. 40). Other proposals do not seem to be moving towards more primitive concepts, for example Putnam's early example of LEMON: natural kind word; yellow peel, tart taste, etc. (Putnam 1970).

Of course, any theory of concepts has to admit of conceptual primitives in the theory (or to embrace an unattractive circularity). That is not peculiar to prototype theories. The problem is, rather, that prototype theories do not seem to move in the direction of more primitive concepts.

Rosch (1978) raises three types of cases in which she worries that prototypes may not bottom out in what she calls 'real world' attributes. First, there are concepts whose attributes seem not to be meaningful without first knowing that the object falls under the category in question. Rosch's example is SEAT as an attribute of CHAIR. The second is the use of attributes such as LARGE which depend upon the category or some superordinate category for their context. For example, PIANO has LARGE as an attribute, but pianos are large in the context of furniture, not buildings. Yet PIANO is a basic level concept, which is supposed to be employed in categorising something as *furniture*. It is not problematic that prototypes contain attributes whose values are contextually determined, but this is a problem if it leads into tight circles in the prototypes. Rosch's third example is TABLE which she found, on asking experimental subjects, has YOU EAT ON IT as an attribute. But the latter surely is a step away from conceptual primitiveness, requiring concepts founded in the human cultural system. It remains at best an open question, then, whether prototype theorists can make good on the claim that their theories move towards concepts which are plausibly primitive.

This line of enquiry points to a further question, not peculiar to prototype theories: how is the content of these primitive concepts determined? Perhaps prototype theorists can rely on very simple prototypes – the prototype for ROUND is: attribute: shape, value: round (assuming ROUND is a primitive concept, in the foregoing sense). But can this be the basis of an adequate theory of content? What falls in the extension of ROUND? Prototype theorists are torn between two options. Prototype theorists might argue that anything the thinker judges to be round counts as falling under ROUND. That sort of verificationism about content is generally unpalatable, but maybe it is acceptable for primitive concepts. Alternatively, the content of primitive concepts may be determined in some other fashion (pick your preferred theory of content). Why then does that theory not apply to lexical concepts? Some of the initial motivation for the prototype as content-determining is thereby removed.

Of course, even if you think that prototypes only have a role to play as a means of identification, you still need an account of how the concepts mentioned in the prototype achieve their content, so you still need a primitive level of concepts. The advantage is that, since concepts do not have their content determined by their prototypes, it is plausible that many lexical concepts are primitive. Prototypes may be employed as a means of identification, but the same univocal theory of content applies both to the lexical concept, and to any concepts employed in a particular identification. It also allows lexical concepts to acquire additional means of identification without thereby becoming a different concept. A thinker may start with a primitive concept like DOG, with a rough ability to tell dogs from non-dogs which does not employ any further conceptual resources. But experience may then allow the development of a prototype for DOG, which adds to the means of identification available to the thinker. In that case, DOG moves from being a primitive concept to one where further concepts are sometimes employed in identifications, but does not change its identity, with its content determined in the same way both before and after the prototype is formed. The prototype theorist has the prototype as part of the reference-determining structure of the concept. So if a new prototype is associated with a concept, its content may change. Indeed, if the content of primitive concepts is not given by a prototype theory, then the whole basis for determining content changes in our DOG example.

The problems with regress and circularity are avoided by one variety of prototype theory, which has prototypes as stored perceptual exemplars. On this view, the prototype does not employ any further concepts; rather it consists of some form of non-conceptual

representation at a relatively perceptual level (which may be a remembered experience, or some form of construction from perceptual experience), together with a similarity metric. None of this need use further concepts. So, it is an attraction of the perceptual-exemplar variety of prototype theory that it is not susceptible to objections of circularity or regress.

6.2 Prototypes of 'well-defined' concepts

The experimental results of Armstrong, Gleitman et al (1983) throw further doubt on prototype theories as an account of conceptual structure. They found that subjects displayed typicality effects even for such well-defined concepts as ODD NUMBER. Concepts in this category have two characteristics. First, subjects agree that membership of the category is an all or nothing matter. Second, subjects agree that whether an instance satisfies the concept is determined by whether it falls under some definition associated with the concept (eg, NOT DIVISIBLE BY TWO). However, subjects consistently rated instances for typicality (3 is more typical than 313), and speed of categorisation correlated with those ratings.

The first lesson to draw from these results is that prototypical structure does not imply that membership of a category is graded, as at first thought. So although prototypes may explain why some concepts have graded application, graded membership need not be a consequence of having prototypical structure. However, Armstrong, Gleitman et al drew a further conclusion from the presence of typicality effects for well-defined concepts. They thought that this shows concepts have a dual structure, divided into a systematic categorical core and an identification procedure, with prototypes forming part of the identification procedure and thus producing the typicality effects which they observed.

That move supports my view that typicality effects are produced by the mechanism for identifying whether an instance falls under a given concept. Armstrong, Gleitman et al appear to agree that the prototype plays no role in determining whether or not a given instance in fact falls under the concept. A concept's extension is determined in another way, in their case by a classical definition. If so, in my terminology, the prototype is not part of the reference-determining structure of the concept. Notice that where the content of the concept is determined by classical definition, as is plausibly the case with ODD NUMBER, the concept is in fact a complex concept, and so we would expect the prototype to be merely a means of identification, for the reasons discussed in section (5) above.

6.3 Ignorance and error

What Laurence & Margolis (1999) call 'the problem of ignorance and error' arises both for classical and prototype theories of concepts. The difficulty is that a person may operate with a prototype which is (intuitively) wrong, in the sense that it represents as statistically reliable features of a category, features which in fact are not reliable. Or a thinker might be ignorant of several of the features which are part of most people's prototype of the concept. For example, early experience might lead a thinker to have colour: WHITE as a constituent of their prototype of CAT. That prototype would then pick out intuitively the wrong class of objects. Prototype theorists of conceptual content are forced to stick with the conclusion that whatever is picked out by the prototype falls within the extension of the concept. That produces worries about lack of intersubjective comparability and intrasubjective stability of concepts (see next subsection). It also makes it impossible for a subject to have errors in her prototypes. Yet people do make an apparent / real distinction, and alter their judgements of what is typical of a category, without their concept changing.

This all makes sense if prototypes are only part of the means of identification, with extension determined by another means. Then there is scope for altering the prototype in order to make it better at picking out the extension – the subject may be wrong in some of the features represented as part of the prototype. But only in separating content determination from prototypicality is a theorist entitled to the distinction between correct and incorrect features of a prototype (or between instances that appear to fall under a concept and those that do in fact fall under it).

6.4 Psychological Generalisation and Concept Stability

Concepts form part of many psychological explanations: a person acted thus and so in part because he possessed a particular concept. Underpinning such explanations are generalisations about what people do on the basis of particular concepts. As generalisations, they apply across groups of people. If there are to be any useful such generalisations, then, people must share some concepts. Thus, we want to be able to say of two people, that they acted thus and so (eg, in the same way in relation to a given individual) in part because they employed the same concept. That is, a theory of concepts must individuate entities that are plausibly inter-personally stable. Similarly within a given individual: psychological explanation mentioning concepts relies upon the fact that psychological generalisations which mention concepts pick out the same concept on each

occasion of use. So there is a strong argument that much of our psychological explanatory practice, especially that which adverts to propositional attitudes, requires a robust notion of concept identity, both within an individual over time, and between individuals. Prototype theories of concepts threaten to undermine the possibility of such intra- and inter-personal concept stability.[12]

If the content of concepts is determined by their prototypes then the natural place to locate concept identity is in the prototype. However, it is implausible that all the thinkers appearing to use the same concept are actually employing the same prototype. On the face of it, even attaching slightly different weights to the same values of the same attributes (in the style of Smith, Osherson et al 1988) introduces a difference in the prototypes. In addition, it is pretty clear that typicality effects change over time, as people gain experience with the subject matter of the concept. Such changes ramify, since prototypes employ further concepts which are also subject to such changes. If each such change were to change the concept, then there is no prospect of explaining an individual's behaviour in terms of generalisations which mention particular concepts – there is no prospect of concept identity across individual or over time, and there is little prospect of content similarity being stable enough to perform the same role. The sort of holism threatens which is faced by most conceptual role theories of concepts; however, it is even worse, since prototypes are even more finely articulated than some standard conceptual roles.[13]

One line of response would be to place concept identity at the level of reference. So two concepts are identical (tokens of exactly the same conceptual type) just in case they have the same extension, irrespective of whether different prototypes determine this extension in different cases. There are two difficulties with this line of response. The first is that it eliminates concept-based explanations at the neo-Fregean level of sense: the sorts of explanations which explain why thinkers with co-referential concepts may behave differently, but in ways which are predictable and explainable on the basis that they have *different* but co-referential concepts (Hesperus / Phosphorus cases, etc.). The second

---

[12] I use 'stability' to be neutral between identity and similarity. If concept identity is rarely realised, then there must be a robust theory of content similarity which underpins the necessary generalisations.

[13] The difference is one of degree, since conceptual roles are often individuated at the level of entire belief systems (e.g, Block 1986). The point is that the machinery of prototypes – weights, similarity metrics, etc. – produce even more loci at which people can differ in their belief systems, and so make identity of prototypes an even less realisable ideal than identity of conceptual roles.

difficulty is that identity at the level of extension will not give the necessary concept stability, if extensions are determined by the prototype. Even small changes to the prototype will produce small changes to the extension of the concept, and so block concept identity, even if such changes in extension are irrelevant to all uses of the concept in practice.

A more promising line of response is to formulate some notion of 'similar enough' prototypes to count as identical concepts. This is standard for conceptual role theorists.[14] The idea would be to carve up conceptual space as determined by prototypes at roughly the level of sense, so that people with radically different prototypes have different concepts but people with roughly similar prototypes count as having the same concept. However, prototype theories have as yet produced no proposals of this type; and perhaps for good reason, since it is not at all clear how the appropriate similarity relation could be drawn. The holism mentioned above is never far away: how can two concepts be similar if, for each, every concept in their chains of conceptual inter-relationships affects their identity?

Admittedly, the difficulties of intra- and inter-personal stability are faced by many theories of concepts, so it is not a powerful objection that prototype theories face them too. However, it is a relevant concern in the project of building an adequate theory of concepts. If prototypes are relegated to an identificatory role, then there is some prospect that the way content is determined will allow a practicable notion of concept identity to be formulated. If prototypes are part of a concept's reference-determining structure, then intra- and inter-personal stability are very substantial difficulties that prototype theorists have still to overcome.

(7) TYPICALITY EFFECTS WITHOUT PROTOTYPES

The empirical psychological work carried out in the context of prototype theories of concepts is one of the great success stories in the study of cognition. Stable typicality effects have been discovered which are confirmed across a wide range of experimental paradigms: typicality judgements, feature lists, speed and accuracy of categorisation, etc. This has been important and fruitful work, producing valuable results for the theorist to work with. A recurrent danger in the study of concepts is that the philosopher should

[14] E.g., Segal (2000).

construct an elaborate theory based on intuitions plus little empirical evidence. Such theories have a poor prospect of hitting the truth. The need to account for typicality effects provides a welcome constraint on theorising about the kinds of things that concepts might be.

However, it would be wrong to conclude that these typicality effects should be incorporated into a theory of concepts by identifying concepts with prototypes, or claiming that prototypes are responsible for determining the content of concepts. I have discussed objections to prototype theories of concepts which stem from considerations of: (1) circularity, (2) well-defined concepts, (3) ignorance/error about prototypes, (4) stabilising functions, and (5) conceptual combination. None is a conclusive objection. But together they amount to a strong argument that prototypes do not form part of the reference-determining structure of concepts – that in virtue of which the content of a given concept is determined.[15] Prototypes are better viewed as being part of a categorisation mechanism: the process by which thinkers in fact carry out categorisations.

The empirical typicality effects can be adequately accounted for if prototypes are means of categorisation, without being content-determining. That interpretation also allows that experience of a category can alter its prototype, without introducing conceptual instability. On this view, it is no surprise that the prototypes of complex concepts are not directly related to those of their constituents – that they are sometimes absent and often informed by experience of the complex category.

So, prototypes are not content-determining. Furthermore, they may not even be psychologically real, in some cases. Connectionist models show how typicality effects can arise in the absence of any kind of prototype structure. It is not novel to observe that connectionist networks sometimes show typicality effects.[16] The distinctive contribution of my clustering proposal is to show how such effects arise. The explanation makes clear that no prototype need be employed.

In a connectionist network, categorisation consists in activation arising within a cluster in state space. That may be an output cluster, if the network has been trained to perform such a categorisation. Or, it could be a cluster in hidden layer state space, arising from performing some other task. Subsection 7.2 of chapter 2 showed how such hidden

---

15 Laurence & Margolis (1999) reach the same conclusion.

16 Rumelhart, Smolensky, McClelland and Hinton (1986).

layer clusters could become available for use in their own right.[17] Section (1) of the current chapter explained how clusters with complete contents could nevertheless form a constituent of or input to a conceptual system.[18] When activation arises within state space, it may fall near the central tendency of some cluster, or towards its margins, or in no cluster at all. Prototypical samples will produce activation in the centre of some cluster. This will be a measure of typicality in the context of the set of samples on which the network was trained. It may not capture what is typical about the real instances that do fall in the extension of the referent of the cluster (as ascribed by 3.5.1). Nonetheless, it will give rise to prototypicality effects.

The main typicality effect that clusters explain is accuracy of categorisation. Recall that training samples cause a layer of the network to differentiate into clusters. The training samples that fall within a particular cluster are then considered. The content ascribed to a cluster is some property common to and distinctive of the samples producing activation within it (recall ch. 2, 3.5.1). Suppose the network is presented with a new sample which also has this property (but differs from the training set). The network may or may not categorise it correctly. If the new sample produces activation within the pre-existing contentful cluster, then the network is classifying correctly; otherwise, it is not. The distribution of training samples in a cluster produces a typicality space. New samples are more likely to be correctly classified if they are close to the prototypical training sample, and less likely to be correctly classified if they are atypical with respect to the prototypical training sample.

Notice that there are two elements to this account: typicality spaces formed by training samples; and the similarity space of the hidden layer within which clusters lie. That is, there are two sorts of similarity in play: similarities between samples, and similarities between patterns of activation that they produce. These must be kept distinct, since they play different explanatory roles. The typicality of a new sample is measured by its similarity to the range of training samples that fall under the same category. Whether it is correctly classified depends upon it producing activation in the cluster that represents that category. That will depend upon where in state space the new sample produces activation. Since typicality with respect to training samples in a given category correlates with centrality in the cluster which represents that category, typicality

---

[17] See also ch. 3, ss. 6.2, especially the account of competitive networks.

[18] See also ch. 3, ss. 2.1.

can predict correct classification (since it can predict whether the new sample will fall within the relevant cluster). Therefore, clusters can give rise to and explain accuracy of categorisation.[19]

In static networks, the only typicality effect predicted or explained by the model is accuracy: that accuracy of categorisation should correlate with typicality. However, dynamic networks can account for a further typicality effect: that speed of categorisation correlates with typicality. Typical samples will give rise to activation of one of the attractor processes in a dynamic network, allowing it to settle quickly into its end state. Un-typical samples will produce initial activation further away from the system's component attractor processes, and so the system will take longer to arrive at a classification. Thus, prototypicality will predict speed of categorisation in a dynamic system.[20]

In rejecting prototypes as content-determining generally, I leave open the question of what does determine conceptual content. What the connectionist model shows is that alternative theories of content are compatible with the existence of typicality effects, even in the absence of prototype structures.

A further suggestion from the empirical data is that there is a basic level of categories, being the most general level at which instances share a relatively large number of features. Connectionist models throw some light on the existence of a basic level. As explained in chapter 2 (subsection 7.4), state space admits of clusters of clusters, and sub-clusters within clusters. So the topography of state space can, in some cases, be attributed content at more than one grain of analysis. That can account for a hierarchy of categories, reflected in a hierarchy of concepts. What of the privileged basic level? Perhaps such a level can be divined, as that corresponding to the way that clusters are best individuated, from the point of view of understanding the operation of the system.[21] The superordinate and subordinate clustering could then be seen as subsidiary phenomena. On this account the basic level would not be objective: it would be a side effect of the type of task on which the network was trained, and the type of samples presented to it in training. In other words, human basic level concepts are only basic from the point of view of the kinds of task in which they arise and are employed. That is perfectly plausible.

---

19 See also ch. 2, ss. 7.3.

20 See also ch. 2, ss. 7.3.

21 And, if appropriate, taking account of processing topography analysis, as suggested in ch. 2, ss. 5.4.

Overall, however, my speculations about a basic level in the clusters in state space are only offered tentatively.

In any event, I do not claim that connectionist networks can model all the varieties of typicality effects. Many such effects doubtless do arise from inter-relations between concepts. Such structure falls outside my connectionist approach.[22] The value of the connectionist model is that it shows how all the pieces can come together, albeit with limited application: there is no internal prototype, the representational structure that does exist does not determine reference, but that structure does produce and explain some typicality effects. Consequently, there is no temptation to the view that reference-determining structure also gives rise to prototypicality effects. Furthermore, this model is offered in a context where there is a positive alternative theory of how the content of such representations is determined.

The other virtue of this discussion of prototype theories is that it arrived independently at the conclusion that structure of representational space need not be content determining. That rebuts any suggestion that that idea is a problematic feature of my connectionist theory of content in chapter 2.

[22] See ch. 3, sec. (2).

5

Externalist Syntax?

(1) EXTERNALISM

1.1 Taking an Interest in Syntax

Most philosophical theories of content take for granted the existence of syntax. It is assumed that a theory of content is only concerned with showing how the content of pre-existing syntactic vehicles is fixed. Amongst the properties that provide a basis for content determination are properties of the syntactic vehicles. Such properties are presupposed. Individuation of syntax in realistic systems is a question to be answered by brain scientists: difficult, messy, and philosophically unimportant. However, there are good reasons to think that the problems with individuating syntax in realistic cognitive systems go beyond the merely practical. I take a close look at the issues. I will argue that considerations about appropriate contents can have a role to play in narrowing down candidate syntactic mechanisms. So theories of content should not take syntax for granted. This chapter will show that it is permissible for contentful considerations to play a role in characterising syntax. I explore the issue by investigating the recent suggestion that syntax could be externalist.[1]

Chapter 2 propounded a theory of syntax and content for connectionist systems. According to that theory, items from outside the system do have a role to play in

[1] Bontley (1998).

characterising the system's syntax: clusters are formed out of activation patterns in the hidden layer produced by the set of *samples* on which the network was trained. That makes it look as if the syntax of connectionist system could be externalist. This chapter examines what that claim could amount to, and elucidates various strengths of the claim. It goes on to consider whether syntax in general could be externalist and, if so, in what sense. In the process it dispels any worries about the role of external samples in my theory of syntax for connectionist systems. It also concludes that various strong versions of externalism about syntax are untenable.

1.2 Wide and Narrow Psychology

Semantic externalism began as a thesis about the meanings of words, and was soon extended to claims about the contents of thoughts. That spawned a vigorous debate about whether the explanatory properties relied upon by psychology are or must be externalist.[2] One way of addressing that question is to analyse a particular psychological theory, which is what a number of commentators did in respect of Marr's (1982) theory of vision.[3] One claim sometimes made in that debate was that Marr's theory is individualistic because it individuates psychological states syntactically. Bontley (1998) countered this move by pointing out that syntactic states might also be externally individuated. But what would that mean?

1.3 Externalist Syntax?

Bontley claims that some theorists advocate wide psychology on the basis of two premises: (i) contents are causally explanatory and (ii) contents are externalist; and that others argue for narrow psychology by denying premise (i), contending that syntax does all the causal work. As Bontley observes, the latter argument presupposes that syntactic states are individuated individualistically.

Bontley argues for externalist individuation of syntactic entities on the basis that they arise from a functional characterisation of the internal organisation of a system; and that here function should be given an aetiological-teleological interpretation, as it is in

---

2 Having changed his mind on a key issue, Fodor can illustrate both sides of the debate: Fodor (1980) and Fodor (1994).

3 Burge (1986), Segal (1989), Davies (1991), Segal (1991), Egan (1991), Egan (1992).

teleosemantic theories of content. Thus, the very individuation of the vehicles of content depends upon how the implementing mechanism is supposed to operate, where purposes are explained in evolutionary terms. It would follow that syntactic states are individuated in part by their extrinsic properties.

1.4 Tying Down the Possibilities

Having motivated a discussion about syntactic externalism, I now need to be clearer about the various types of claim that might be made. Externalism is most often formulated as a thesis about supervenience. Consider some cognitive system S (person, animal or computer). Externalism about syntax is the claim that the syntactic properties of S do not supervene on its intrinsic properties.[4] The idea of supervenience is that the supervening properties (here, syntactic) do not distinguish between any entities that cannot already be distinguished by the properties on which they supervene (here, intrinsic properties).[5] On standard views, syntax arises out of the kinds of causal properties of a system that are described by physics (which are assumed to be intrinsic). The possibility under consideration here is that, in some cases, the complete set of physical, intrinsic properties of a system does not fix its syntax.

What are the particulars of which those properties are predicated? I will sketch three options, before defining each carefully in the following paragraph. The first option is to hold the view that, although each of the entities which together make up a system can be picked out in terms of intrinsic properties, a correct syntactic description of those entities depends upon factors outside the system. A second option is more strongly externalist: claiming that the very way the system is divided up into entities of which syntactic properties are predicated depends upon external factors. On that view, it would

---

[4] Strictly, 'intrinsic properties shared by duplicates', since there are plausibly some intrinsic properties not shared by duplicates, e.g., of system S, the property of being identical to S.

[5] Formulation taken from Davidson (1993). More carefully:

> A set of mental properties **M** *supervenes* on a set of physical properties **P**
>
> iff (by definition)
>
> necessarily$_1$ {for any x and **M**-property M, x has M at time t only if there exists a **P**-property P (physical base property) such that x has P at t, and necessarily$_2$ (anything that has P at a time has M at that time)}.
>
> (Taken from Kim 1998, p. 9, and Sturgeon 1998.)

be wrong to think that the same internal entities had different syntactic properties in different environments, since the very entities would change depending upon the environment. The only entity comparable across environments, so that we could ask about different syntactic properties of *it*, would be the entire system S. Bontley's view falls in this category, because the very syntactic items are individuated teleologically. Finally, a third option is open, which is even more radical. That would be to hold that the entities which have syntactic properties extend outside the system into its environment. In which case, it is not even clear that we are dealing with the same system if we change the external-world context.

To tie matters down, I assume that the entities which realise syntactic types are spatio-temporal particulars: states, processes, events, physical grounds for dispositions, etc. That assumption should be uncontentious given the naturalistic context of the debate. Then the following are the three possible strengths of externalism about syntax:-

1.4.1 The entities within a system of which syntactic properties are predicable are spatio-temporally local and do not vary depending upon external factors. However, the syntactic properties which those entities instantiate do not supervene on the intrinsic properties of the system (i.e., the way to classify those entities together as different physical instantiations of the same syntactic type depends upon factors external to the system).

1.4.2 The entities within a system of which syntactic properties are predicable are spatio-temporally local, but the way to divide up a system into entities of which syntactic properties are predicable does not supervene on the intrinsic properties of the system. A fortiori, the syntactic properties instantiated by entities within the system do not supervene on the intrinsic properties of the system. (I.e., the very entities in a system which can be syntactic vary depending upon external factors).

1.4.3 The entities associated with a system of which syntactic properties are predicable are spatio-temporally extended, and the way to divide up a system and its environment into entities of which syntactic properties are

predicable does not supervene on the intrinsic properties of the system.[6] A fortiori, the syntactic properties instantiated by the system in its environment do not supervene on the intrinsic properties of the system. (I.e., the very entities which can be syntactic extend outside the system.)

'Spatio-temporally local' here is intended to be neutral between the entities being:-[7]

(i) the entire system (i.e., the entity is spatio-temporally co-located with the system);

(ii) temporal stages through which the entire system passes (i.e., the entities are spatially co-located with the entire system); and

(iii) proper parts of the system (i.e., the entities are spatially located within the entire system).[8]

(2) SYNTAX

2.1 Classical Computationalism

The syntactic states of a classical computational system are realised locally, typically in the electrical current in some electronic circuit or the stored charge in magnetic or other media. The syntactic states are specified by the system's designers. In virtue of that design, syntactic states can be used to realise computational programs (by moving from a machine code that connects more or less directly with the basic syntax, through higher-level programming languages, to some specification of the computational process in algorithmic terms). When the program is taken to be performing some function, the

---

6 As will be seen, I find it hard to make sense of this claim, but for the formulation here to work 'system' has to be thought of as specifying something smaller than the syntactic items with which it is associated. 'System' should be taken to be the supervenience base of intrinsic properties (which do not determine the syntax) which are shared between the cases in which 'external' factors lead to change in the syntactic entities.

7 According to the usual understanding, syntactic items fall into category (iii).

8 Where the syntactic entities are proper parts of the system, a very strong internalism could hold that their syntactic properties supervene on the intrinsic physical properties *of those entities*. That is implausible, however, because on any view the syntactic characterisation of a particular entity will depend upon its interrelations with other entities in the same system.

syntactic states can be interpreted as bearing the contents attributed to them by the program. However, processing in the system is insensitive to content, and proceeds solely in virtue of syntactic properties. The trick of programming is to ensure the syntactic items are chosen so as to respect their intended contents.

It has proven very fruitful to use this model as a source of intuitions about cognition. But it does also produce difficulties. One is to drive worries about what explanatory role content could possibly have: syntax seems to be doing all the causal work.[9] For the purpose of this chapter, I want to point to an even more basic problem with using the classical computer as a model for human cognition: it leads to the presupposition that the vehicles of content can be taken as given by a philosophical theory of content.

Although the computer model is a source for presupposing an uncomplicated internalism about syntax, it also militates against a too-strong internalism. Syntactic states are proper parts of the system. Thus, a very strong internalism could hold that their syntactic properties supervene on the intrinsic physical properties *of those realising entities*. That is implausible, however, because on any view the syntactic characterisation of a state of a classical computer will depend upon its interrelations with other syntactic states in the same computer. So, syntactic properties supervene on the intrinsic properties of the whole system, but not upon the intrinsic properties of the individual entities which realise those syntactic states.

That observation provides a model which might apply more widely. We ask about the syntactic properties of some spatio-temporally local particular within a system, and we find that we have to look past that particular to the system as a whole to answer the question. Perhaps in some cases we will also have to look outside the system, notwithstanding the fact that the entities in question are spatio-temporally local particulars intrinsic to the system.

2.2 Realistic Candidates for Cognitive Systems

In humans, and other realistic candidate cognitive systems, the syntax is much less evident than in a classical computer. Most theories of content assume that the vehicles of content are given: they are sitting there, ready-individuated, bearing all their properties, waiting for a theory of content to come along and tell them what they mean. So theories of

[9] An early and well-formulated expression of this concern is Field (1978).

content take for granted an entitlement to rely on all the properties of a candidate vehicle without further explanation: its causal relations with the outside world, its causal interrelations with other internal states, perhaps even its evolutionary history. That presupposition is theoretically unsubstantiated. It may be supported in part by tacit reliance on the model of classical computation.

The human brain is so little understood that it is not yet known what its content bearing states are. Two ideas dominate, each underwritten by a different experimental paradigm.[10] Single unit recording in live animal brains (and occasionally human brains) drives the idea that single neurons represent, and do so by means of an elevated firing rate. There are difficulties even here, since most neurons have a base-level firing rate, so to interpret a neuron as either on or off requires a threshold to be specified and motivated. Furthermore, some single unit studies suggest that variations in the pattern of firing carry representational content. For example, within a particular brain area, the phase difference between a single cell's firing pattern and a cross-population 'theta rhythm' is thought to code for direction.[11] So, even if it is assumed that single neurons represent, it is still unclear what the basic syntactic items are. And great controversy abounds about whether all, most or no representation is achieved by single neurons (the so-called 'grandmother neuron' debate). Theorists who use neuropsychological evidence, brain imaging (PET and fMRI) or studies of event-related evoked potentials (EEG and MEG) are sometimes localist, but can also allow that whole regions of the brain are implicated in representing each of a whole class of stimuli. The model is of representations of a certain type distributed in some region of the brain, with different patterns of activity in that region representing different things. There are severe methodological difficulties with trying to design investigations to arbitrate this debate,[12] but studies continually appear in support of the localist interpretation[13] and, more rarely, explicitly in support of the distributed view.[14]

---

10 See also ch. 3, ss. 6.1.

11 O'Keefe & Burgess (1996), Maguire et al (1998), Burgess & O'Keefe (2001).

12 Cohen & Tong (2001).

13 E.g., Downing et al (2001).

14 E.g., Haxby et al (2001).

It might be assumed that these are just epistemic problems: that we don't yet know enough about the brain to be able to divide up its workings into syntactic units. However, the current state of play suggests a deeper problem than that. The way the details of brain mechanism are realised are quite well understood, including the structural and chemical changes that cause a neuron to fire, and the mechanisms of chemical neuromodulation; thus explaining a lot about how patterns of firing spread. Much is also known about the brain's anatomical connectivity. The problem of syntax may be deeper than an issue of practical complexity. There are many different ways of describing what goes on in brains, from the molecular upwards, each of which contributes to a story about how the causal processes unfold; but it is not clear, even in principle, how to divide these processes up into the units which are relevant from the point of view of the brain's computational processes. If that is indeed the situation, then an understanding of the processes at a contentful level might usefully contribute to that characterisation. The contentful understanding can derive from many sources: from the resources of everyday psychology, through the careful experimental studies of cognitive psychology, to the mathematical formalisms of decision theory. It is hard to deny that an understanding of the nature of human cognitive computation in contentful terms should contribute to the practical task of identifying the vehicles of contents inside people's bodies and brains. I claim that it cannot be excluded that this connection exists as a matter of theoretical principle, too.

Modellers building connectionist systems similarly assumed that it was unproblematic to identify their syntax, guessing that the basic syntactic items were single node activations, and that each pattern of activation was a different syntactic item. Chapter 2 shows why that was mistaken. Taking each potential pattern of activation as a syntactic item is rather like taking every different voltage level in some semi-conductor circuit as a different syntactic item. Both are fallacious. In the computer case, it is easy to see why, because we know how syntax was designed into the system. In the connectionist case, since its end-state performance was not designed, but rather developed in the course of supervised learning, it is not so clear what the syntax should be. I suggest it was the model of classical computation that led to the assumption that the vehicles of content in connectionist systems are readily identifiable. It may similarly be a mistake for theories of content to take for granted, in real-world cognitive systems, the availability of a pre-existing syntax from which the theory of content is independent, and the properties which it can presuppose.

2.3 Extended Cognition

There is a movement amongst some cognitive scientists away from aiming to program intelligence into systems that compute over internal representations. Their motivation derives partly from frustration with classical artificial intelligence research, and is partly inspired by the multifarious clever ways in which living organisms take advantage of their ecological surroundings to support complex behaviour. The 'extended cognition' programme designs real moving systems, consisting of sensors and motors as well as microprocessors, which can perform some specified complex task (e.g., Brooks 1991). The possibility arises in such cases that the vehicles of representational content might themselves spread outside the system that we would normally take to be the agent. I am not here considering cases where some cognitive process includes off-loading representational content into the world as a stage of cognitive processing: doing logic using written symbols, storing information in writing, working out an argument by writing it down. Even if these are examples of cognitive processes, with some of the cognitive steps taking place outside the thinker, they are not worrying. They are cases where the external symbols are fully external – written words, etc. – not ones where we are tempted to say that a single representational vehicle is partly inside and partly outside the thinker. Cognitive loops, stages of which involve external symbols, seem to me unproblematic. Contentful vehicles which straddle thinker and environment are much more so. Representationalism seeks to explain cognitive processes in terms of causal interactions between successive contentful vehicles. Currently, there is no clear model of how such interactions could take place between extended entities that straddle the boundary between agent and world. The integrity of a syntactic item as a single unit derives in part from the fact that causal processes act on it in a unitary way – to explain the processing it undergoes, the theorist does not have to advert to things that happen separately to different parts of the representation. The theory can treat the representation as a whole, and explain how the various stages of processing apply to it. It is far from clear that entities that crossed the agent-world boundary could be syntactic in this sense.

Many working in this field agree with me. They accept the phenomena of extended interactive systems performing intelligent tasks, but deny that they can be explained in representational terms. They rely upon these examples to illustrate 'causal spread': that brain, body and world all make contributions to adaptive success, so that none can be a preferred locus at which to locate representations. Since it would be excessively liberal to

suppose that there are representations in every part of these extended processes, commentators like Webb conclude that none of these areas contain representations.[15] Some claim that all intelligent behaviour can be achieved in analogous ways, and so cast wider doubts on representationalism. On the other hand, Andy Clark has suggested that representational content may still be attributed to some extended cognitive systems.[16]

Most radically interpreted, the phenomena of 'extended cognition' motivate a very strong externalism about syntax, with spatio-temporally extended syntactic entities – where specification of the basic entities, as well as the syntactic properties to be attributed to them, depends upon factors outside the system (i.e., the variety of externalist syntax defined at 1.4.3 above). However, it is very hard to see what role the syntax is playing, if that were the case (on which, more in section (4) below). A more moderate position accepts that, when the interacting entities whose properties explain some complex behaviour extend significantly beyond the system driving that behaviour, representational attribution ceases to be available to explain the behaviour.

Wheeler & Clark reject the claim that one can justify a general anti-representationalism on the strength of causal spread alone, although they do think that some such considerations in particular systems can count against representational explanation:

> 'What all this shows, we think, is that there is some plausibility to the claim that certain forms of large-scale causal spread threaten to undermine representational forms of explanation in cognitive science.'
> (Wheeler & Clark 1999, p. 111)

Even if these are not cases of representation at all, the possibility of an extended basis for intelligent action creates a further difficulty for the assumption that the syntax of real-world systems may be taken as given in specifying a theory of content.

---

15 Webb (1994).

16 His excellent philosophical summary of the then state of play is Clark (1997).

2.4 A Suggestion

In subsections 2.1 to 2.3 I have argued that there are some genuine problems with presupposing a syntactic characterisation of real-world cognitive systems. To proceed from here, I will take a closer look at the role of syntax in my connectionist theory of content (section (3) below). It will provide an alternative model to the classical computational picture, which will help make the issues clearer. Then I will argue for a positive account of the role that syntax has to play in a theory of content, and show how it seriously limits the extent to which syntax could be externalist (section (4) below). But first I will sketch a suggestion about how syntax works. By seeing my goal in advance it should be easier to follow the ensuing discussion.

So here, roughly, are my answers to three questions about syntax. (a) What does a theory of content add to a syntactic understanding of the operation of a system? It tells something about the behaviour of the system in relation to its environment – provides a way of describing its behaviour which connects with the system's worldly context, for the purposes of predicting and explaining what it will do in such contexts.[17] (b) What then does syntax add to a contentful explanation? It explains how those contents are physically realised. The presence of a physically-realised syntactic typing licenses a different pattern of contentful prediction and explanation than that available to a theory which is purely behaviourist or attributionist about content. The presence of a syntactic characterisation underpins what is useful about representationalism over behaviourism. (c) So what does syntax add to a mechanistic description of the implementational mechanism? My answer is roughly as follows: there are many different ways of describing an implementing mechanism, not all of which line up with the vehicles of content. A syntactic characterisation divides up the operations of the implementing mechanism at a level of generality and abstraction that fits with the contentful story.

My answers to questions (a) and (b) are fairly standard. It is my answer to question (c) that I will substantiate and defend in what follows. The theme will be that syntax must be about implementing mechanisms, but characterised at a level that fits with content. Contents invoke a form of explanation that goes beyond implementing mechanism. But representationalism only gets a grip when there *is* an implementing mechanism. Syntax is the way of specifying that mechanism that also connects with the content-involving story.

---

[17] This question is considered in more detail in ch. 6, pt. II below.

(3) THE CONNECTIONIST CASE STUDY

A first step towards substantiating my idea about syntax in general is to see it illustrated in connectionist systems. My theory of syntax and content in connectionist systems (chapter 2) gives a role to external-world samples in individuating syntactic states. The proposal was to divide points in activation state space into clusters based on their proximity. (The same points carry through *mutatis mutandis* in respect of processes in dynamic networks.) The criterion of proximity is internally specified, supervening only on properties of the system itself. However, to answer the question which regions of proximity should be taken to be syntactic items, the theory looks outside the system: it relies on the existence of some set of external samples on which the network was trained, in respect of which it performs correctly. An array of points in state space is generated with respect to that set of samples. Syntactic items are not arbitrary regions of proximal state space, but only clusters in the array of points that correspond to training samples.[18] Once individuated, these clusters can be described in purely intrinsic terms; and they are not extended entities, but entities within a layer of the system. Each token state which falls under a particular syntactic type is just a particular pattern of activation across a given layer. The role of external world samples is to collect these entities into the relevant syntactic types.

How does this connectionist syntax measure up against the range of externalist positions set out in subsection 1.4? It illustrates the weakest of the externalist theses entertained: 1.4.1. Syntax is realised by patterns of activation, which are spatio-temporally local: they lie within the connectionist network.[19] Those entities are the same irrespective of external factors. However, the way to classify those entities together as different physical instantiations of the same syntactic type depends upon factors external to the system. That is, the syntactic property which a pattern of activation instantiates does not supervene on the intrinsic properties of the system.

The metric of similarity between patterns of activation is fixed internally: it is proximity in state space. But the question of how to use that internally specified similarity

---

[18] Recall that points represent levels of activation in a layer (usually hidden) produced in response to samples. So a point 'corresponds' to a training sample when it is a point in state space that represents the activation pattern produced in a layer of the network in response to that sample.

[19] Processes in dynamic networks are also spatio-temporally local, so the point carries over to my theory of content for dynamic networks.

metric to sort entities into types depends upon the external samples. It is that sorting which is relevant to the system's operation in the environment in which it was trained. Since the similarity metric is internally specifiable, once a syntax has been arrived at, it can be fully described in internalist terms. That reflects the fact that it is a genuinely causal-mechanistic way of describing the system. But from the purely internal perspective a whole plethora of such mechanistic characterisations are possible, so external factors are needed to tell which is the syntax.

Although my connectionist case study does illustrate a very moderate externalism about syntax, it does not support the stronger positions defined in 4.1.2 or 4.1.3 above. The entities which fall under syntactic types are patterns of activation, and do not vary depending upon external factors (only the syntactic properties of these particulars do vary). A fortiori, the vehicles of content are not spatially extended outside the connectionist system.

(4) FINDING A ROLE FOR SYNTAX

4.1 Syntax Characterised by its Theoretical Motivation

There are many reasons in support of the idea that mental states have contents: their power in predicting and explaining behaviour, their occurrence in the everyday practices of folk psychology, their causal efficacy, etc. We seem to have a pretty good grasp of how to attribute and use content, and these attributionist practices are refined, by study and analysis, into a science of rationality.[20] Having explained the connectionist case, I can re-address the question: what does syntax add? That is, what is the theoretical motivation for the claim that contentful states are realised by particular physical states that can be syntactically characterised? (It is rather ironic that this question should arise, given that the traditional problematic – which reflects the presupposition of syntax – concerns the converse direction, and asks what content adds to syntax, Field 1978.)

[20] I have in mind something like the attributionist 'Unified Theory' advocated by Davidson (1995), consisting of a theory of meaning in Davidson's Tarski-like form, together with a rational decision theory based on attributing subjective probabilities (degrees of belief) and preferences evaluated on an interval scale (strengths of desire). Theories of content based only on attribution are found in Ryle (1949), Dennett (1987) and Davidson (1984).

To expand on the rough sketch given earlier (subsection 2.4), my answer derives from the motivations for rejecting behaviourism. Behaviourism resisted contentful explanations of human and animal behaviour, and certainly rejected the idea that there are internal vehicles of content. The 'cognitive revolution' made it respectable to rely upon internal representational states as part of an explanation of behaviour.[21] The assumption that contents are realised by real, individuable, internal states leads to a different range of predictions, and different constraints on explanation, from a purely behaviourist or attributionist approach to content. I call this move away from behaviourism 'representationalism'. Representationalism requires that there is a syntax – some way of individuating vehicles of content within the internal mechanism. So, what syntax adds to the purely attributionist employment of content ascriptions are the benefits of representationalism over behaviourism.[22]

The main motivation for representationalism is to account for the causal efficacy and intra-personal stability of contentful states. Contentful states should be physically realised in an agent if they are to be causally relevant to the agent's behaviour. Furthermore, the causal effects of a state with a given content are relatively stable over time, allowing contents to figure in psychological generalisations. That stability is explicable if the same syntactic item carries the content on each occasion. That is, intra-personal stability is explained by the vehicle of content being the same (from the point of view of internal mechanism) each time the content is a cause or effect. Internal states with causal powers can similarly take part in the inferential processes which form part of psychological explanation. These benefits of representationalism require something more than the bare existence of a physical realiser for every representational state. They accrue because there is some property of the mechanism of the realising system which is common between the various vehicles with the same content. Some stable property of the implementing mechanism is needed if the fact of physical realisation is going to account for: causal efficacy, a role in psychological generalisation and intra-personal stability over time. Syntax plays that role.

In short, rejecting behaviourism motivates the existence of syntax, where a syntactic characterisation is a matter of internal mechanism. Representational content

---

21 Chomsky (1959).

22 I will not here argue for representationalism over behaviourism – that debate has already been exhaustively investigated – but just elaborate the motivations for the move to internal vehicles of content.

gets its grip just when contents which would be licensed by a purely attributive approach map onto a characterisation of the system's internal mechanism. The mechanistic description which allows for that mapping is the system's syntax. So, I conclude that syntax must be a way of characterising a system's mechanism of operation.

That conclusion is obviously compatible with syntactic internalism. Does it allow any variety of syntactic externalism? The three possibilities are set out in subsection 1.4. Of course, it is a matter for empirical investigation to discover what the syntax of cognitive mechanisms is. However, 1.4.3 is excluded, given the discussion of what syntax must be for a representationalist: context-dependent extended vehicles are unlikely to figure in any characterisation of a cognitive system's internal mechanism of operation. That leaves the other two options: entities internal to the system whose syntactic properties (1.4.1), or whose very individuation so as to be suitable for bearing syntactic properties (1.4.2) depends upon external factors. Which way to go will depend upon future research about the implementing mechanisms of cognition, and upon how those empirical discoveries interrelate with the various theoretical reasons in favour of representationalism which were mentioned above.

The spirit of representationalism favours 1.4.1: not only are all the realisers of a given syntactic type internal entities within the system, there is also something intrinsic in common between those that fall under that syntactic type. Put another way, a purely intrinsic description of the syntactic mechanism of operation will be available. The role of external factors is to choose amongst many possible intrinsic characterisations of the mechanism, settling as syntactic that mechanistic characterisation which aligns with content attribution, and hence relying on external factors in taking that step. The idea is that all mechanistic characterisations are based on intrinsic properties. Amongst those one is identified as syntactic in virtue of external factors.

However, for the reasons given in the next subsection, I cannot demonstratively exclude the idea of 1.4.2: that intrinsic factors do not determine how the states of the system should be divided into entities of which syntactic properties may significantly be predicated.

4.2 The Possibility of a Teleofunctional Mechanism

Recall that Bontley (1998) envisioned syntax to be externalist because it is fixed by a system's internal functional organisation,[23] where functions are to be understood in evolutionary-teleological terms.[24] There has been extensive debate about how biological functions should be ascribed.[25] On the simplest reading, some entity is picked out by its current intrinsic properties (the heart is picked out as a lump of flesh) and we ask of it what functions it serves as a matter of evolutionary purpose. On that interpretation we again have only the very weak externalism of 1.4.1. Bontley himself accepts that a teleofunctional approach to syntax should be substantially different to teleofunctional accounts of content, if both are to play separate explanatory roles.[26] The moderate externalism of 1.4.1 would differentiate syntax from content, since teleo-*semantics* clearly allows much stronger brands of externalism. So, even a teleofunctional account of syntax might only lead to the moderate externalism of 1.4.1.

However, these are deep waters, dependent in part on progress in biology in carrying out functional explanation, so I cannot rule out the possibility of 1.4.2 – that the very division of a system into entities for the purpose of teleofunctional explanation depends upon aetiological factors.

4.3 A Moderately Externalist Syntax

The previous subsection gave a reason why 1.4.2 cannot be categorically excluded. However, there are several considerations which suggest that 1.4.1 is much more plausible. The motivations for representationalism sit more naturally with 1.4.1, making less sense if the division of a system into mechanistic entities were to depend on historical factors.[27] Stability over time, and consistency of inferential connections both push in the

---

23 Bontley calls these 'narrow' functions, in a usage which differs from the usual application of the labels 'wide' and 'narrow' in the debates about externalism, since Bontley's 'narrow' functions are nonetheless externally individuated.

24 Subsection 1.3 above.

25 See, for example, Millikan (2002), and the references therein.

26 Bontley (1998), p. 572.

27 Subsection 4.1.

direction of syntactic types having intrinsic physical similarities between successive realisations.[28] For these reasons, in what follows I will assume only the very moderate externalism defined in 1.4.1. But this should be read with the caveat that 1.4.2 is not demonstratively excluded.

4.4 Application of the Conclusion

What does this moderately externalist syntax look like in practice? I have already given two examples. The first is in my theory of content for connectionist systems. There, states which fall under syntactic types are patterns of activation, irrespective of the appropriate syntax. External samples are used to divide these patterns of activation into clusters in state space.[29] Since the metric of similarity can be specified in internal terms (proximity in state space), once a syntactic characterisation has been arrived at, it can be fully described in internal terms. The mechanism of operation could be described in terms of architecture, weight matrix and activation function; but that fails to connect with content at all (see chapter 2). A more abstract description in terms of regions of state space can connect with what the system is doing in contentful terms. However, there are indefinitely many ways to use the metric of state space proximity to divide up state space into regions of contiguity. Each is a mechanistic characterisation. Only one connects with content attribution. That one gives the syntax. Since content attribution is used to select amongst candidates for syntax, the syntax of a system is not fixed by its intrinsic properties (moderate externalism 1.4.1).

The second illustration of moderately externalist syntax is a teleofunctional account of syntactic mechanism, whatever the representational system. The internal mechanism of a system can be causally described in several ways, but only one connects with its evolutionary purposes. So teleofunctions would decide that only one of the potential ways of describing a representational system's internal mechanism counts as syntactic.

There may be an interesting overlap between these two cases. The clusters that develop in a connectionist system as a result of training can be ascribed teleofunctions.

---

[28] Although, suitable stability of the environment over time, on which external factors depend, would also allow for the requisite stability.

[29] Or attractors / principal component processes, in a dynamic network.

That function derives from the purposes for which the system was trained. That is, the interpretation given to the outputs, as a goal of training, sets up a relational proper function for the hidden layer(s).[30] Clusters which develop during training acquire a derived proper function as a result, deriving from the specified relational function. That makes a nice connection with my connectionist theory of content: of the various ways of describing a connectionist system's mechanism of operation, the description which is syntactic is the one which derives from the system's input-output purpose, and which is the basis for a contentful explanation of what the system is doing. That connection between content and developmentally derived functions is suggestive. It will be explored in chapter 6 below.

Finally, moderate syntactic externalism might arise in many other ways. A correct theory of content in some domain might involve external factors in a variety of ways: causal-informational connections, output-success connections, etc. Any such external constraints on content may have a role to play in individuating syntax. The basic idea is that many mechanistic explanations of the internal operation of the representational system are available: the one which is syntactic will be fixed so as to connect with the external factors upon which the particular theory of content depends.

(5) CONCLUSION

The ascription of representational content to explain the behaviour of a system only gets its distinctive explanatory grip when the theory holds that those contents are realised by the system's syntactic mechanism of operation. Which means that syntax must be basically internalist. At the very least, the vehicles of content must be spatio-temporally internal to the system. Most likely, there must also be intrinsic similarities between the various entities which, within a given system, realise a particular syntactic type. However, none of these considerations excludes a role for external factors in choosing amongst various true descriptions of a system's mechanism of operation, so as to single out one such mechanism as providing a syntax.

One application of this idea has syntactic types delineated in terms of an internal similarity metric (one that depends only on intrinsic properties), but with external factors playing a role in setting the borders between similarity classes.

---

[30] Using the terminology of Millikan (1984).

In short, the type of moderate externalism about syntax presupposed by my theory of connectionist content is unobjectionable. Representational explanation relies on an intersection between world-involving contents and internal mechanisms. Syntax lies at that intersection. So, the question of which internal mechanism is genuinely syntactic may depend upon some of the external factors upon which content ascription depends. But only in moderation.

6

Content Determined Partly by

Ontogenetic Factors

I. Could Ontogenetic Factors Play a Role?

(1) INTRODUCTION

An unorthodox feature of my theory of content for connectionist systems in chapter 2 is that it assigns a role to developmental factors in content determination. Recall that, in a feedforward multi-layer connectionist network, only after training does the state space of a hidden layer differentiate into clusters. So the system's syntax only arises as a result of development. And, to individuate syntax, the actual samples on which the network was trained are employed.[1] Clusters are not a feature of the network alone, nor of how activity of merely possible inputs would be distributed in state space. They are individuated by plotting the points in state space corresponding to activity produced only by the samples actually used in training.[2] Thus, if the system had developed in different circumstances, it would have a different syntax. That is for two reasons. Firstly, because

---

1 Ch. 2, ss. 3.1 & 3.5.

2 Strictly: those amongst the training samples that lead to correct responses once the network has reached criterion, which will be almost all of the training samples.

different training samples would produce a different end-state weight matrix.[3] Secondly, because different samples would then be used to individuate the clusters.

Developmental circumstances play a second role in the theory of chapter 2. They contribute towards the ascription of content. Recall that the content of a cluster depends upon the output task for which the network was trained.[4] Properties tracked by a hidden layer must be relevant to the output task. Thus, the output task for which the network was actually trained will affect the content to be attributed to its internal states.

In the present chapter, I will argue that theories of content more generally may have to allow a role for ontogenetic factors in content determination. My claim is that appropriate theories of content for other systems may also entail that content is partly determined by the individual circumstances of a system's development. If so, this feature of the connectionist theory is not unorthodox.

I do not argue that it is an adequacy constraint on any theory of content in any domain that the determination of content should partly depend on developmental circumstances. My only claim is that this is an open theoretical possibility in at least some domains. The question of which theory of content is correct has a long history of protracted investigation. Even now, no theory is without difficulties. Nor does any theory command a consensus among philosophers or psychologists. There is not even a most-favoured candidate leading the field. Accordingly, new avenues of enquiry offer a fruitful means of increasing our understanding of the issues. Ontogenetic factors offer one such line of enquiry. Historical factors are not new in theories of content. Both causal and teleosemantic theories of content rely upon them. I'm suggesting a different type of historical factor: the particular environmental circumstances that caused a given individual to develop that specific representation. My connectionist case study shows that it may be worth considering ontogenetic factors. The current chapter argues that they could be relevant, more generally, to content determination. My aim is to show that the idea is plausible enough to merit further investigation.

Developmental factors are historical. There is a general objection, much discussed in the literature, to placing any reliance on historical factors in a theory of content. The objection is that it follows from such reliance that a molecular duplicate of a person with contentful states which arose at random – a swampman – would not have contentful

---

[3] In empirical studies with different training samples the weight matrices at least look completely different.

[4] Ch. 2, ss. 3.5.1.

states. That is claimed to be an intuitively unattractive consequence of any theory that relies upon history. I take swampman, not as a thought experiment valuable for the intuitions it reveals, but as an arresting way to pose a general question: why should historical factors be better than current ones in fixing content? To answer that question properly, one needs a clear idea of what contentful explanation is for. Thus, in part II below I explore some motivations for employing representational explanations. I then go on in part III to use these motivations to answer the swampman-type objection. My conclusion will be that there is no decisive obstacle to relying upon historical factors as partly determinative of content, either in the connectionist theory in chapter 2, or more generally.

The chapter is structured as follows. Part I (i.e., sections (2) to (6)) makes the case that developmental factors could play a role in content determination. Section (2) uses some examples to promote a role for circumstances of development. I consider the intuitive content of representations arising from various learning mechanisms, both at a relatively low-level (2.1), and in distinctively human cognition (2.2). Section (3) endorses Laurence & Margolis' (2002) recent contention that it is a substantial constraint on an adequate theory of content that it be compatible with a credible account of the psychological development of the representations to which it ascribes content. I argue that this constraint suggests a stronger conclusion: that content is directly determined, in part, by such developmental circumstances. Section (4) examines the gap left by teleosemantics for developmental circumstances. I show that, as a result, teleosemantic theories of content are committed to individual developmental circumstances as partly determinative of content. Section (5) concerns innateness. It is theoretically difficult to draw a distinction between the innate and developmental contributions to the nature of psychological mechanisms. It may even be impossible. That is another reason why ontogenetic factors are important, if phylogenetic factors are: because they cannot effectively be separated. Section (6) tentatively suggests how the developmentally-based part of a theory of content might look, and how it might connect with the distinctive nature of content-based explanation.

Part II (i.e., sections (7) to (10)) carries out the intermediate task of characterising what content attribution is for. It starts by explaining the question, in section (7). Theories of content often aim just to give naturalistic co-extension conditions for content that get the right answers in intuitive cases (ss. 7.1). The resulting theories

nevertheless contain implicit answers to the question; some address it explicitly (ss. 7.2). In section (8), I explore three types of answer to the question. Section (9) draws a moral from that discussion, namely that both inputs to and outputs from a system should play a content-determining role. That was a feature of my theory of connectionist content in chapter 2. Section (10) addresses a question that is closely related to the issue of the purpose of content attribution, that is, are contents causally efficacious? In that section, I sketch two possible theoretical positions that can be taken on the issue.

Finally, in part III the foregoing discussion is used to answer the swampman challenge. Section (11) asks why current factors should be inadequate to fix content. Swampman illustrates the claim that current properties are sufficient for content. Section (12) dismisses some responses to the swampman case that miss the point, or are addressed to the details of the thought experiment, rather than to the underlying issue. In section (13), I pose the challenge specifically in relation to my theory of connectionist content, and answer it. Section (14) canvasses three answers to swampman offered by teleosemantic theorists. Section (15) draws out two considerations that lie behind all these answers and contends that one of them, at least, supports reliance on historical factors as content determining. Section (16) concludes that, as a result, swampman is not an objection to the suggestion made in the current chapter – that content may be partly determined by individual circumstances of development.

(2) EXAMPLES FROM HUMANS AND OTHER ANIMALS

2.1 Low-level Learning

This subsection gives examples of four learning systems in animals where, intuitively, the end state representation refers to the thing encountered during the development of that end state. These mechanisms are also likely found in relatively low-level human psychology.

The first is imprinting. That is the process in which a newly born animal learns to behave in a special way towards a parent: to follow it around, demand food, etc. Lorenz famously demonstrated the phenomenon by leaving his rubber boots for young geese to see as they hatched. They would then faithfully follow him around the town.[5] The circumstances in which imprinting will occur, and its behavioural consequences, have been

[5] Rose (1992), p. 58.

extensively described in chicks (Bateson 1966). The mechanism seems to give rise to a new representation: the chick comes to identify and keep track of something new, and behave in various ways in relation to it. The object first presented is clearly part of the cause of this representational development. The representation also seems, intuitively, to refer to that object: it is supposed to keep track of the object first seen.[6]

A second example is provided by the cognitive maps that some animals develop as a result of experiencing a local environment (Pearce 1997, pp. 203-214). For example, rats can learn the layout of a maze of platforms hidden underwater, or an array of objects hidden around a room. There is good evidence that this representation is stored in so-called 'place cells' in the hippocampus (O'Keefe & Nadel 1978). The new representation is caused by the spatial layout of the environment in which it developed; and, intuitively, represents it. A particular rat's cognitive map seems to be about its learning environment, and not other places that happen to have the same geography, or in which the rat's map-guided actions would turn out to be successful.

A third illustration is aversion learning. This is the striking phenomenon in which an animal will avoid a food if the taste of it is followed by sickness.[7] As in classical conditioning, an UC, sickness, comes to be associated with a CS, the taste. However, unlike classical conditioning, the learning occurs after only one trial, and the aversive stimulus need not be paired in time with the taste, but may occur several hours later. The substance with that taste is part of the cause of the new aversion.[8] And the new disposition is, intuitively, an aversion to that substance.[9] To test this, consider an animal with its taste buds subsequently reversed by some physiological re-wiring. The animal would then avoid the wrong things – that is, its aversive representation continues to refer to that which originally caused it.

Finally, consider a regular case of classical conditioning: learning to identify a foodstuff by sight as well as taste. In primates, it seems this is achieved in part by the

---

6 Those in a theoretical frame of mind might dispute this. Doesn't the representation refer to the chick's mother, whatever it was hapless enough actually to imprint on? My use of the example relies on a more naïve intuition.

7 Shepherd (1994), p. 633-634.

8 If the sickness were paired with no CS, then no new aversion would arise.

9 Whether it is an aversion to things with that taste, or things of the same kind as the cause of the aversion, does not matter for the example.

development of neurons with finer sensitivity in the orbitofrontal cortex (Rolls & Treves 1998, pp. 155-159). The new representation is caused to develop by the foodstuff with that smell and taste. Plausibly, the referent of the representation is that foodstuff.[10]

None of these examples is revolutionary. Doubtless, many other theories could make good claims for rival content assignments. However, the examples have a common thread, which suggests a special role for an ontogenetic factor, both as the causal source of a new representation, and as its referent. More modestly, they illustrate that it is at least plausible that the circumstances in which a representation developed constrain the content that is to be ascribed to it.

2.2 Human Learning

There is strong evidence that humans have a specialised capacity for recognising faces.[11] The first indications came from the existence of patients with a selective deficit in the ability to recognise faces, called prosopagnosia (Sacks 1985). There are now several converging lines of evidence that face recognition is performed by a dedicated system in the brain. Neuropsychological studies show that damage to a specific brain area is associated with severe prosopagnosia. That has been confirmed with functional imaging, and by electrical measurement and stimulation inside the brains of epileptic patients.[12] The area specific to faces is called the fusiform face area, located near the junction of the occipital and temporal lobes of the cerebral cortex (although many other brain areas are also involved in processing faces, including prefrontal areas). On experiencing a novel face a person develops the ability to produce a new representation (which is at least partly located in the fusiform face area, and is distributed across that area), which she employs in recognising that face in the future.

---

[10] This example is more controversial. Perhaps the animal has an existing representation of that foodstuff, and has simply learnt to distinguish it in a greater variety of circumstances. That interpretation is resisted if several different foods share the associated taste, since the new representation will be specifically sensitive to the food with the relevant appearance. Even so, this is a case where different theoretical perspectives will motivate different content assignments. It is less clear here that one option is more intuitive than all the others.

[11] Kanwisher (2000).

[12] Cohen & Tong (2001) summarises the evidence.

It seems obvious that this mechanism's function is to enable people to recognise each other by their faces. It is part of the way that humans keep track of conspecific individuals. So the representation refers to an individual: the person who caused that representation to develop (call him 'S', for source). A different individual, experienced in unusual visual conditions, could later cause the same representation to be tokened. It would then misrepresent (*that is S* would be false). Similarly with look-alikes. We use face recognition to build up a body of knowledge about how we should act towards a person, and about what he will do. It would be a mistake to project these expectations across to a different individual who happened to look very similar. It is not superficial similarity that grounds the projection of attributes from occasion to occasion. It is the fact of encountering the same individual on each occasion (since many attributes of an individual person are stable over time). And the source of that mistake would be a false representation. The error would start when seeing the look-alike and thinking *that is S*. The content of that thought is false because the face-tracking representation refers to the original individual, and not anyone else.

Similar considerations have been used in the broader context of the philosophy of language to argue that causal history partly determines the content of proper names (Kripke 1972). However, the conclusion is more compelling in the case of face recognition, both because the phenomenon is simpler and better-described, and because the correct answer is more obvious. The evidence is overwhelming that the ability to recognise a particular face arises only as a result of experience, and is implemented by means of an internal representation. It is then hard to resist the conclusion that the particular circumstances in which one of those abilities develops – the person you see when you learn to recognise someone new – partly determine the content of the resultant representation.

The same thing occurs in higher level cognitive systems. Since these systems are less well described and understood, the content ascription is correspondingly more contentious. I will take as an example the acquisition of concepts of natural kinds. There is good evidence that the way children categorise changes dramatically as they grow up. Even when newborn they can keep track of objects, by trajectory and number.[13] Ingenious experiments based on violation of expectancy[14] show that babies soon come to

[13] Carey & Xu (2001).

[14] This is operationalised as looking time, graded from videos by naïve independent observers. Some critics object to the assumption that increased looking time implies violation of expectancy. However, what is

differentiate solid objects from portions of stuffs,[15] and then begin to track objects by category (e.g., animate vs. inanimate) until they can eventually differentiate objects at the level of natural kind terms: by species, etc.[16] By the age of 2-3 years children can categorise a wide range of objects on the basis of what they look like and what they do: their characteristic features. But then there is a dramatic shift. Children stop relying upon a wide range of characteristic features and shift to a smaller core of 'defining' features as the basis for their category judgments (Keil 1989). This shows up in overt category judgements, and in the range of new exemplars to which children will project existing known properties. It is also found implicitly in the way that children project what they learn about things one can do with members of the category.[17] By 4-5 years old children are very good at penetrating beneath surface appearances (Gelman & Wellman 1991). Their judgements come to be based more on objects' insides[18] or, for animals, their lineage.[19] Most strikingly, this "characteristic to defining shift"[20] is much more pronounced in relation to natural kinds than artefacts.[21] With artefacts, there is a more subtle shift towards greater reliance on an object's function.

These results have been used to argue that children are committed to there being a hidden essence shared by samples of a natural kind.[22] There is also evidence that the 'folkbiological' method of taxonomising local and fauna is culturally universal,[23] although whether essentialist dispositions are universal is controversial.[24] Some philosophers have relied on the existence of essentialist dispositions to argue that the reference of natural

---

important is the existence of statistically significant differences in looking time, demonstrating that the babies differentiate the situations, however we choose to describe it.

15 Soja, Spelke and Carey (1991), Huntley-Fenner, Carey & Solimando (2002).

16 Mandler (1994), Carey & Xu (2001).

17 Mandler (1998).

18 Gelman & Wellman (1991).

19 Keil (1989).

20 Keil & Batterman (1984).

21 Keil (1989).

22 Keil (1989), Gelman & Wellman (1991), Gelman & Coley (1991), Gelman, Coley & Gottfried (1994).

23 Atran (1990), (1999).

24 Gelman & Hirschfield (1999), Atran (in preparation).

kind concepts is fixed externalistically, so as to depend upon the kind which, in the thinker's actual environment, gives rise to the relevant surface properties.[25] This has led Segal (2003) to deny the essentialism,[26] along with the externalist model. However, for my purposes the answer to this empirical debate is not crucial. It does not matter whether or not concept users are in fact committed to natural kinds having some hidden unifying essence, since I reject the idea that a thinker's conceptions determine the reference of her concepts in these kind of cases.

For my purposes, the importance of the developmental studies is to show that an explanation of children's deployment of concepts must advert to more than surface appearances (irrespective of whether children have essentialist beliefs or not). If a theorist is to explain the patterns of behaviour of older children and adults, she cannot base her explanation only upon the ways that objects appear. As the experiments show, it is the reidentification of something underlying that explains how the subject will act on a new instance, and which properties they will project to it. What is the referent of such a concept? Let's answer that by asking what it takes for a subject to be getting it right when he uses the concept in relation to a new instance. The answer is that he must be right that the new object has the property which he projects to it, or affords the action that he performs on it. For such projections to be justified, there must be something in virtue of which the instance shares those properties with the original samples that he learnt about. Notice that, to be useful, the property / affordance projected must go beyond the way that the new instance is identified as falling under the category. Suppose you had to check than a fruit was red, round, crisp and tasty before classifying it as an apple. Then inferring from *that's an apple* to *that's tasty*, while justified, would not tell you anything new. So the relevant underlying feature must give rise to both the properties used to identify instances and to the non-apparent properties that can thereby usefully be projected. So here is the picture: concepts of natural kinds are employed to project useful properties and affordances from learning samples to novel instances. For that to work, novel

---

[25] Putnam (1970), Rips (1989); Laurence & Margolis (2002) also rely upon psychological essentialism.

[26] For more detailed argument, see Segal (forthcoming). Strevens (2000) also expresses dissent, but his objection is to reliance, for example, on an essence of tigerhood, in addition to the property of being a tiger. However, this misinterprets the developmentalists' claims. They are only committed to children tracking the property of being a tiger, distinct from the surface attributes of tigers. They are not committed to the existence of both a property of being a tiger and an essence of tigerhood. An essence is nothing more than that which gives the property of being a tiger its unity.

instances must fall within the same category as the learning samples, where membership of that category is the causal source both of the properties the thinker relies upon to identify an instance as falling under the concept, and of the properties a thinker thereby projects to those new instances.[27]

I rely on this picture of the operation of natural kind concepts for two conclusions. Firstly, the reference of such concepts depends upon the useful functions which they perform. *A fortiori*, it is not determined by the beliefs thinkers associate with the concept (their conceptions), essentialist or otherwise. Secondly, reference depends upon the samples the thinker experienced when he originally developed a concept for the category. The reference is some feature of *those samples* which allows him to project knowledge about the original samples to new instances. Given original samples of a different kind, but with the same surface features, the causal basis for the projection of properties would be different, so the referent would be different. When he uses the new representation in respect of an instance of a different category (e.g., he thinks *that is an apple, eat it*, on seeing a wax apple), then the error consists in identifying the wrong thing – something that does not share a projective ground with the original learning samples. The fact that this new use is a misrepresentation shows that the natural kind concept is tied to the learning samples. Thus, its referent depends in part upon the particular circumstances in which the concept developed.

This picture of natural kind concepts emerges from the developmental studies, and what they show about the function of such concepts. However, it has very interesting parallels with my theory of content in connectionist systems (chapter 2). I argued there that contents should be ascribed to hidden layer clusters when a network manages to project correct classificatory practice to new samples, and when those samples lie outside the training set but fall into existing clusters. In that case, the network is keeping track of some underlying feature shared by training samples and new samples. That explanation ascribes content to the clusters, the contents being those shared features. Training samples cause the cluster to develop. Property projection outside the training set calls for a contentful explanation. And the content to be ascribed is thereby determined partly by the training samples and training task. Similarly with natural kind concepts. Experience with initial samples is the developmental cause of a new concept. Empirical work shows

[27] These properties need not apply to all category members, but only to arise reliably enough from category membership to be useful.

that such concepts are used to project properties to instances that do not share surface appearances with the training set. So the subjects are keeping track of some underlying feature shared by training samples and new instances. That explanation ascribes content to a concept, the content being that shared feature. Initial experience with samples of the kind causes a new concept to develop. The concept is the basis for projecting properties to samples outside the training set. And the content to be ascribed to the concept is thereby determined partly by the training samples.

The picture I have painted is closely related to Millikan's (2000) theory of substance concepts. In particular, I draw from her the idea that use of these sorts of concepts depends upon projecting learned properties to novel instances. That entails that members of the category share some underlying ground that is the causal source of the projected properties. It is these underlying grounds which Millikan calls 'substances': they are the causal source of the co-projection of a variety of properties over instances. However, I rely on developmental considerations more explicitly than Millikan does. I use the idea that a new substance concept will develop as a result of experience of samples of the substance. Then the reference of the concept will depend upon the identity of those learning samples. Millikan can allow something similar. Her substance concepts are abilities to identify substances. The reference of the concept is given by its natural purpose: the function of the ability is to identify some particular substance, and that substance is the referent of the concept. Millikan's natural purposes are given by natural selection. However, most identification abilities have not evolved directly, but are produced in the course of experience by relational mechanisms which have evolved to produce such abilities.[28] Their functions derive from the function of the learning mechanism. The function of the learning mechanism is relational: to produce new abilities that function thus and so.[29] The new abilities so produced derive their function thereby. Strictly, Millikan does not need to specify the learning mechanisms. All that she needs is that there are learning mechanisms that allow abilities to identify substances to be

---

[28] The theory of relational and derived functions is explained further in section (4) below. For more detail, see Millikan (1984, pp. 39-50) and (2001).

[29] If it is not obvious from Millikan (2000) that abilities to recognise substances have derived proper functions, Millikan says (personal communication):

> "The mechanism that recognises any particular substance will not be just a general purpose mechanism in any case but a special purpose mechanism produced by a learning process governed by a general purpose mechanism operating in a particular situation."

acquired. However, in my view this lacuna should be filled. Content is determined by the function of the learning mechanism. Therefore, if Millikan is to make a convincing case for the contents she claims, she should specify the learning mechanisms that give rise to such contents. She is reluctant to make any detailed claim which could be hostage to empirical disconfirmation. However, for my purposes, only a minor extension is required. I add the idea that, when a general learning mechanism operates in a particular situation to produce a new substance concept, features of that situation determine the function of that ability, and hence the content of the concept. That is perfectly compatible with Millikan's theory of relational and derived functions, and may even follow from it. If so, Millikan's theory of substance concepts also supports my claim that the samples which are experienced when a new natural kind concept develops partly constrain the content of that concept.

To recap this subsection, I have given two examples of relatively high-level human representational abilities: face recognition and natural kind concepts. In respect of both, there is a good argument that a correct theory of content for such representations will show that their content is determined partly by the circumstances in which they developed.

(3) A Theory of Content Must Be Compatible With Representational Development

Laurence & Margolis (2002) have recently argued for a weaker claim: that an adequate theory of the content of mental representations must be compatible with plausible accounts of how those mental representations arose in psychological development. Their project is to reject Fodor's strong concept nativism. They start by re-construing Fodor's nativism as a challenge: how can primitive (i.e., unstructured) representations be learned? Lacking an answer, Fodor concludes they must be innate. Laurence & Margolis disagree. They argue that there are plausible theories of the acquisition of new primitive representations; i.e., accounts that do not require the new representations to be structured out of existing ones. They take the learning of new natural kind concepts as an example, and work through an empirically justified account of their acquisition.[30]

[30] Their account relies upon the kind of evidence mentioned in the previous subsection, so I largely agree with it. However, they suggest that natural kind concepts require essentialist conceptions. They need the essentialism because of their commitment to Fodor's asymmetric dependence theory of content. The essentialist disposition makes it the case that causal relations between non-referents and the concept are asymmetrically dependent on the causal relation between the referent and the concept. I disagree with

The challenge is to fit the development with the theory of content. That theory will spell out the factors which determine content. Laurence & Margolis take these to be informational connections (they work with Fodor's asymmetric dependence theory of content). The developmental account must show how the end state comes to display the appropriate features, so that content is appropriately determined by the theory of content. The challenge is to demonstrate compatibility between means of acquisition and the theory of content. It is not simply that it would be nice to have an account of how the representational states are acquired. The constraint is stronger. An adequate theory of content must be compatible with the appropriate content-determining factors being acquirable, according to plausible accounts of development, based on the best empirical evidence.

Of course, one way that the theory of content could be compatible with the developmental story is if developmental circumstances partly determine content. That is my claim. The thrust of Laurence & Margolis' argument comes close to that stronger suggestion:

> 'For the present purposes, however, the crucial point we want to emphasize is … that questions about the nature of concepts are intimately bound up with questions about how they are acquired.'
> 'So even with primitive concepts, an investigation into how they are acquired seems likely to say quite a lot about their nature.'
> (Laurence & Margolis 2002, both at p. 50.)

I agree that the nature of representations is intimately bound up with how they are acquired. That intimacy, I suggest, may be reflected in their contents, such that a representation would not have the content it does if it had not been acquired in the circumstances it was.

Where I disagree with Laurence & Margolis, however, is with their assumption that the development of syntactic items is less problematic. They assume that potential vehicles of content are available, the properties of which can be adjusted in content-relevant ways, so that a vehicle comes to have the features which determine its content

---

the chosen theory of content, the reliance on thinkers' conceptions as content-determining and the resultant view that essentialism is indispensable.

appropriately (in Laurence & Margolis' case, being the appropriate informational relations). Thus, as part of their account of the acquisition of natural kind concepts, they say:

> 'She sees a new object that has features that suggest that it is a natural object of some sort. … upon encountering the item, the child *releases a new mental representation* and begins accumulating information about the object and linking this to the representation.'
> (Laurence & Margolis 2002, p. 42, italics added.)

More likely, part of the process of developing a new concept is to develop a new syntactic item which can be the vehicle of that content. Laurence & Margolis agree that the representation has to develop properties appropriate to its content. What they miss is that this very process may be what differentiates the representation into a new syntactic type. (Chapter 2 shows that this is what happens in connectionist systems; chapter 5 demonstrates that the model could have general application.) Thus, I argue, there are good reasons to add to the scope of Laurence & Margolis' claim. Not only must a theory of content be consistent with a semantic account of representational development. It must also be consistent with a syntactic account of representational development. Indeed, the two may be inseparable. Together, they furnish a substantive constraint on an adequate theory of content.

(4) DEVELOPMENTAL FACTORS IN TELEOSEMANTICS

We saw in subsection 2.2 above that Millikan's theory of substance concepts has content fixed by natural function. At the end of that subsection, I argued that for acquired concepts, that function will also depend upon the particular circumstances in which the concept developed. Evolution only fixes the relational function of a learning mechanism. That is true of teleosemantic theories of content in general.

There are a number of different teleosemantic theories, but they share a central idea. A representational system has the function of ensuring the output of some second, co-operating system coincides with some condition in the environment. By ensuring, in the evolutionary past, that the output coincides with that environmental condition, the representational system has helped predecessors to survive and reproduce. The content of a representation is that environmental condition with which it is designed to make the

output of the co-operating system coincide.[31] Paradigm examples are bee dances, beaver splashes and frogs' tongue-dart reflexes. The evolved purpose of each is rather determinate. Consequently, representational content is fixed by those purposes. However, more needs to be said about the application of natural selection to intentionality. That is because it is possible for many organisms to produce entirely new representations, never seen before in the history of that organism — representing something that neither they, nor their ancestors, have ever encountered before. How can these novel items have their content fixed by a theory that relies upon evolutionary functions?

Face recognition provides an example (see the start of subsection 2.2 above). The evolutionary function of the face recognition system is to produce new subsystems. Each new subsystem has the function of recognising a particular individual's face. Although a particular individual could be entirely novel, unlike any other person in the history of the species, a representation of her is produced in the normal way by a mechanism that does have a history and an evolutionary function — the evolutionary function of producing sub-mechanisms that recognise individuals by their face. The new sub-mechanism derives a function from the evolutionary function of the learning mechanism that produced it, together with the particular circumstances in which it was formed.

That is a powerful idea: items without an evolutionary history may nevertheless have a biological function, which derives from the evolutionary function of some mechanism selected in the past to produce sub-mechanisms of the same type. In Millikan's terminology, the evolved mechanism has a *relational* function, and the new product of that mechanism a *derived* function.[32] These terms apply even to non-representational mechanisms. Millikan uses the example of the chameleon's skin. The mechanism that makes a chameleon change colour has a relational evolutionary function: to produce a skin colour that matches the chameleon's background. It is by performing exactly this function in the past that the mechanism has contributed to reproduction of the species, and hence reproduction of the mechanism itself. But a particular shade of skin colour adopted on a particular occasion may be entirely new in the history of the species, perhaps because an individual has strayed into a new environment. Nevertheless, this new shade has a function deriving from the function of the relational mechanism. Schematically:-

---

[31] I return to teleosemantics in ch. 6, ss. 8.2.

[32] Millikan (1984), ch. 2.

Relational function:

to match whatever the background looks like at the time

Particular background on a given occasion:

red and green polka dots

Novel derived function of the skin colour on that occasion:

to match red and green polka dots

Relational and derived functions are essential to account for the capacity to represent something never before encountered in the history of the individual or of the species. Where we can ascribe to a device a quite specific relational evolutionary function, then that relational function will be very informative about the function of the derived mechanisms. So, with the face recognition mechanism, the relational function licenses the attribution of a quite specific content to the resultant representations: they each represent the face of some individual. Where the relational function is more abstract, it will have correlatively less to say about the content of a particular representation. More of the content determination must advert to the individual circumstances under which the representation was learned. In such cases, intentionality still derives ultimately from evolutionary functions, but the content of a particular representation is fixed much more by the circumstances of individual learning history than by the evolutionary history of the species. Thus, teleological theories of content allow a substantial role for the circumstances of individual development in content determination. The nature of that role will vary, depending upon the learning mechanism concerned. As yet, teleosemantic theorists have said little about the different sorts of dependence on developmental environment that will follow from different kinds of representational development.

(5) DIFFICULTY OF THE INNATENESS CONCEPT

So far in this chapter, I have been taking for granted a distinction between innate and acquired mechanisms. Thus, in the last section, evolutionary functions and the circumstances of development were given separate roles in content determination. However, those roles may not be so easily separable. Griffiths & Grey (1992) argue that an organism inherits a whole developmental system from its ancestors, including many of the external environmental factors which contribute to its development. There is no sense to

be made of separable contributions from genes and environment, they claim.[33] Certainly, evolved traits needn't be entirely determined by the genes. In fact, on examination, that idea looks like a non-starter. These worries have led to an intense contemporary debate about the innateness concept.[34]

This debate has led to a growing appreciation that the influences of genetic and developmental factors on an end-state trait are intricately intertwined, at best. It has also demonstrated that a trait that has evolved in phylogeny need not be present at birth, nor need it develop only under the influence of internal causes: it may depend importantly on features of the developmental environment, if those features were stably present during evolution of the trait. This all suggests that, if evolved functions are relevant to content determination, then the circumstances of individual development will partly constrain content, since the two are so interdependent.

The strongest conclusion drawn by objectors to the innateness concept is that phylogenetic and ontogenetic factors are inseparable in the contributions they make to the nature of biological traits. This emerges as a consensus position under the slogan 'nature through nurture'.[35] If correct, this position gives stronger support for my conclusion in the previous section: if phylogenetic factors are important to content determination, then so are ontogenetic factors.

(6) A Tentative Suggestion

So far, we have seen several examples where intuitive content ascription seems to depend upon developmental circumstances. When discussing concepts of natural kinds, I gave an argument why end-state contents should be dependent on content in this way. I have also pointed to two further reasons for such dependence, one drawn from Laurence & Margolis (2002), and the other specific to teleosemantics. What will the ontogenetic clause in a theory of content look like? These arguments give some indication of how the dependence will go. But I do not arrive at a fully-fledged theory of content. Nor do I intend to. The

---

33 Cf. Wheeler & Clark (1999), who try to account for a separate informational contribution from the genes towards traits.

34 Fodor (1981), Griffiths & Grey (1992), Elman, Bates et al. (1996), Cowie (1999), Ariew (1999), Samuels (2002).

35 The title of Ridley (2003) – Matt Ridley the science writer, not Mark Ridley, the evolutionary biologist.

examples and arguments point in the developmental direction, but are consistent with a range of options as to the correct theory of content. In the present section I explore a slightly more concrete theory. I offer it as an illustration of what an ontogenetic constraint in a theory of content might look like, rather than as a settled theory in its own right.

In connectionist systems (chapter 2), real world samples cause the development of a network, from a randomly-assigned configuration, to one that has syntactic structure. When the network is viewed purely internally, it takes patterns of activation that show no intrinsic similarity for the network at the input layer, and transforms them through hidden layer clusters into output clusters. By assigning content to hidden layer clusters and output layer clusters, the network's operation can be understood as making contact with things in the real world (in both perception and action). In particular, patterns of activation at the input layer that appear, from the internal perspective, to be unrelated, are shown to have a unity that consists in common properties of the samples that give rise to them. This embedded way of viewing the system's operation is a convenient way of understanding it. Furthermore, it becomes indispensable, if we are to explain how the network manages to project its correct behaviour from the training set to new samples with different input encodings.

So the system carries out an embedded function: acting in response to things in the world and producing actions characterised in terms of properties of those things. It acquires that function under the causal influence of real world samples during development. Various properties of those samples are the causal source of the organisation which is discernable in the developed system. Furthermore, characterising that organisation in terms of those properties (by seeing states of the system as contentful, referring to such properties) allows us to explain how correct performance projects: from samples in response to which the system developed, to entirely new samples.

The general features of this picture are as follows. (i) a potential syntax of the system describes its mechanism of operation in terms of intrinsic properties of the system. (ii) the right syntax fits with properties in the external environment, allowing the system's mechanism of operation to be seen as performing embedded functions – ones where input causes and output actions are characterised in terms of properties of things in the external environment. (iii) those embedded functions were caused to develop by certain properties of things in the external world and, as a result, (iv) seeing the operation of the system in terms of those properties explains why the mechanism extends to new cases. This

characterises contents as simultaneously: mapping onto syntax, figuring in embedded functions, being the causal source of development of these functions, and thus explaining the projection of those functions to new cases.

Very tentatively, this picture suggests that the content of such a representation is: the property which caused it to develop into part of the realisation of an embedded function that extends to novel cases.

Notice that, to arrive at this tentative proposal for content determination, I have relied on some suggestions about the purpose of contentful explanation. The next part of this chapter (sections (7) to (10)) considers some further answers to the question of what contentful explanation is doing. This is a necessary preliminary to rebutting swampman-type objections to reliance on developmental factors (part III).

II. Why Go Representational?

(7) What is a Theory of Content *For*?

7.1 What Realises Intentionality?

Part I of this chapter gave some reasons why the content of a representation may partly depend on the circumstances in which it developed. In the process, a picture emerged of the kind of thing content ascription would be, if such developmental constraints were to arise. The present part presses the same question on other theories of content: what is the theory *for*? If a theory of content is the answer, what is the question? This is an important issue for two reasons. First, to constrain potential theories. Second, to assess whether it is permissible for a theory of content to rely upon historical factors, which is the topic of the final part of this chapter.

A common theoretical motivation for formulating a theory of content is simply to understand intentionality. Explananda, for such theorists, are the everyday phenomena of intentionality. We predict and explain the behaviour of other people by attributing to them beliefs, desires, intentions, etc. Those attitude states are described using propositions: the belief that p. Thus the content, *p*, individuates the psychological state. And these states have features that are rather peculiar in the natural world: they represent or refer, have truth conditions, and so on; that is, they have intentionality. Something similar may be needed to understand the behaviour of some other animals.

How does intentionality arise in the natural world? That is the question to which many theories of content are addressed. The project is to characterise psychological states in other terms so that we can understand in what their intentionality consists. Many theorists are looking for more than informative truths about intentional states. They seek a naturalistic account: an explanation that shows how intentionality can arise from the world as characterised by the natural sciences, free from intentional notions. So, the aim is to say in what content consists, which is metaphysics. An adequate theory will show how the content of a representation is determined by its non-intentional properties. At its strongest, this determination can amount to reduction; more modestly, to supervenience of the intentional on the non-intentional.

This approach does not directly address the question: why go representational? It starts with our commonsense understanding, according to which some psychological states represent, and seeks to explain it. The main task is to show how a distinction between truth and falsity can be grounded non-intentionally. Consequently, the main test of a theory is whether it delivers the contents we expect. Examples are generated where we think we know whether the representation is true or false, and theories are tested to see whether they agree with our intuitions.

The question of what content attribution is for can still be tackled from this perspective. We will see that in subsection 8.1 below. Once a theory has been formulated, and justified by appeal to commonsense psychology, the nature of that theory itself will say something about the nature of contentful explanation.

7.2 Why Attribute Content At All?

An alternative approach is to start by asking what contentful explanation is up to, and to use the range of answers to inform theory building. In practice, a dual attack is likely to be most useful, combining a pragmatic treatment of commonsense examples with theoretical considerations arising from an understanding of the nature of contentful explanation. However, the latter tack is often neglected as a potential source of understanding.

Field (1978) raises the question in the form of a challenge. He argues that it is possible, without using semantic notions, to state the laws by which a system's beliefs and desires evolve as it is subjected to sensory stimulations, and the laws by which those beliefs and desires affect its bodily movements. Field claims that these laws would characterise functionally the syntax of a system and the type-identity conditions for its

states, thus forming a theory which predicts and explains the behaviour of the system. The theory would then have been constructed without attributing representational content to any of the states. The idea is that this purely syntactic theory will fully predict and explain the behaviour of the system in the light of the sensory stimulations which impinge on it. So what further purpose is served by the attribution of representational content?

One response to Field's challenge is to question the possibility of characterising a system's syntactic structure purely internalistically. If the theorist must look outside the system in order even to type-identify its internal states, then it is not so clear that there is available a fully causally explanatory narrow psychology to rival attribution of representational content. According to the theory in chapter 2, to characterise the syntax of connectionist systems the theorist must look outside the system, and describe its response to external samples. In chapter 5, I argued that this moderately externalist syntax is unobjectionable. If that were true more widely, it would partly answer Field's objection, since syntax and semantics are not then rival types of explanation, but require one another. However, Field's challenge could then be reformulated without referring to syntax: why go outside a description of the system's internal mechanism of operation at all? What is gained by relying on the system's relations to the external world (whether in determining syntax or semantics)?

One view is that we re-label internal states with contents in order more easily to understand their interrelations, and the laws that relate them to the outside world.[36] According to this view, everything the system does is caused by, and could be explained by intrinsic properties of the system, but contents are a useful re-description into terms more easily understandable by human users.[37]

What makes contentful explanation more tractable? One answer is that contents are realised in different ways in different people. So the contentful description provides a scheme that can generalise across many people, precisely because it does not condescend to the detail of the internal causal properties of the contentful states. But why should such generalisations exist? In the connectionist case, there was a reason why different systems should have states with the same contents. Content seems to be more than a mode of re-description that happens, accidentally, to apply to a wide range of systems. If

---

[36] Field's (1978) own answer to the question amounts to the claim that internal states are re-labelled with contents to be faithful to the commitment that intentional states should be reliable indicators of the world.

[37] Stich's (1983) view is of this kind.

there is a reason why the contentful mode of explanation is multiply realisable in physical systems, then that reason will show why content ascription adds something to a purely physical-intrinsic characterisation. It will give autonomy to the semantics.

Thus, in tackling the theoretical question – why go representational? – I am taking up Field's challenge: to show what gives content its autonomy.

(8) SOME REASONS TO GO REPRESENTATIONAL

8.1 Embedded Functions

Fully to justify an autonomous role for content would require a fully satisfactory theory of content. That is too ambitious. My aim is to raise some possibilities, each inspired by existing theories of content. Firstly, I will consider the idea that content ascriptions are part of a functionalist specification of the role played by an internal state – a specification of a functional role that that state realises. Where the functionalist specification adverts to things in the environment, then the functional role will be wide. Therefore, the approach holds that contents specify what I will call 'embedded' functions.

In chapter 5, I argued that content ascription must at least depend upon the embedded functions performed by a system. In the current subsection, I will raise some doubts about whether that is sufficient to characterise what it is to be contentful. However, I agree that it is part of the picture: moving outwards from a purely internally-specified characterisation of a system's mechanism of operation is a move in the direction of content ascription. My claim in chapter 5 was that a theory of content does add something to a syntactic understanding of the operation of a system – at the minimum, it provides a way of describing the behaviour of a system which connects with the system's worldly context (chapter 5, subsection 2.4); and it may do more. What then does syntax add to a contentful explanation? The answer in chapter 5 was that it explains how contents are physically realised, and thus underwrites the commitment to realism about representation, by requiring that different token representations with the same content are physically similar (similar with respect to internal processing within the system).

Therefore, embedded functions give one reason to go beyond understanding a system purely in terms of its internal processing. That is part of the motivation for moving to characterising a system in representational terms: content ascriptions are, at least, part of a functionalist characterisation of certain embedded functions realised by the system. However, I will argue in this subsection that this motivation is not distinctive of content

ascription. Accordingly, it may only form part of an account of what content ascription is for. Subsections 8.2 and 8.3 below will suggest two further motivations which could be appropriate supplements.

The idea that contents are ascribed in order to describe a system as realising embedded functions is implicit in several theories of content that do not draw that moral explicitly. Here, I will consider informational theories, and those that rely upon functional/conceptual/inferential roles. The former are a species of the latter, where the roles are specified only in terms of the system's reaction to inputs.

Informational theories base content in correlations between a representation and the things in the environment that cause it to be tokened, or in covariation between some environmental condition and the occasions on which a representation is realised (Dretske 1981, Usher 2001). However, non-representational states of organisms also correlate with things in the outside world (often the correlations are reliable, or important for the organism). So the informational approach fails on its own to say what is special about the representational case. The theories have a correlative difficulty with explaining misrepresentation. Relying upon causal correlation alone would entail that a symbol represents all of its potential causes and so, when actually caused, could not but represent truly. Of course, informational theories are addressed precisely at avoiding these overly verificationist consequences. Fodor's move is to point to a privileged causal correlation: the one upon which all the others counterfactually depend. That correlation is between the representation and its content. For example, COW means *cow* because *cows* cause tokens of COW; and, although COW is sometimes caused by horses on a dark night, the latter depends asymmetrically on the former. The trouble with Fodor's theory is that the required counterfactuals – the asymmetric dependence of one causal connection on another – are almost certainly underwritten by the contents of the representations in question, which are the very things they are supposed to explain. So, Fodor's theory may be true. It may even be informative, as far as it goes, in giving us an alternative characterisation of the phenomenon. But it falls short of naturalising intentionality, and thus does not suggest an answer to the question of why to go representational at all.

Another approach is to use actual causal factors to circumscribe the relevant causal correlations. The idea is to look at the things that have actually caused a representation to be tokened as a matter of causal law. This tactic is used especially to deal with Twin Earth causes; which are exotic things like twin water, that are not found in the thinker's environment but which, if encountered, would cause the representation to be

tokened.[38] This suggests a first way in which contentful connections might differ from other causal correlations. Contentful explanation might give a privileged explanatory role to the things which a person has actually interacted with. (My suggestion in the previous chapter that theories of content might assign a privileged role to properties of the objects a person interacted with in developing a representation relies on a specific kind of causal factor.)

Definitional and prototype theories of concepts also look to the circumstances in which a concept is produced to determine its content. An object falls in the extension of a concept if it has all the features of the definition, or has sufficiently many of the features of the prototype. The idea is not just that these features determine content, but also that tokening representations of the features causes the concept to be tokened (which has the features as constituent parts). Precisely what is entailed about the nature of content depends, of course, upon how the content of primitive representations is taken to arise. But the general picture is of content being a matter of causal or constitutive connections between syntactic states. What is special about the correlations in these systems is that the states in question have compositional structure. However, whether that is true of all representations is tendentious; it would rule out the kinds of systems of which representational explanations were offered in chapters 2, 3, and in section (2) of the current chapter.

Other inferential role theories do not insist on their representations having constituent structure. To be a state with a certain content is just to be related to inputs and outputs, and to be interrelated with other internal states, in a given way.[39] In principle, 'narrow' conceptual roles can be individuated, in which inputs and outputs are specified in terms of states of the system. Narrow conceptual roles do not add anything to syntax. That is taken to be a virtue, but it does not offer an answer to the question of what is special about contentful explanation. Wide conceptual roles go further. They explain the operation of the system as embedded in its environment, by tying content to how states are caused by things in the external environment, and give rise to external results. Like informational theories, they rely upon a system's world-involving relations. Where informational theories rely only on input-factors – how the system is caused to respond to inputs – wide causal roles can be specified both in terms of sensitivity to

---

[38] Putnam (1975) for twin water; Prinz (2002) for causal factors to exclude it from the referent.

[39] Block (1986).

external inputs, and in terms of the external results that actions of the system give rise to (as well as depending on internal interrelations). So we arrive at a more complex characterisation, such that contents arise in a system that behaves in a stable way in some range of environments, with contentful explanation being a way of characterising its embedded functions.

Embedded functions support the idea in the previous subsection that contentful explanations are more tractable, because they generalise across a range of realisations of the contentful states. Thus, the answer to our question offered by these theories is as follows: contents are functionalist labels for the realising states that play the specified wide functional role.

However, we still haven't reached something distinctive of representation. The only claim is that contentful states are functionally characterised: the state such that it is caused by abc and leads to xyz. That may be true. However, many non-intentional properties can also be characterised functionally,[40] and those functions can be wide. Perhaps there is nothing special about contentful roles. It could be that there is a functionalist specification of each contentful psychological state, but no difference in principle between these and functionalist specifications outside the contentful realm. If so, a theory can equally well specify functionally what it is to be a carburettor, and what it is to be a belief that dogs bite. All that makes the latter contentful is that it falls within the functional role characteristic of beliefs that dogs bite,[41] not that it has some special feature characteristic of contentful states.

The mildly externalist syntax discussed in chapter 5 is consistent with this view. Amongst the ways of describing the system's operation, one fits with its realising an embedded function, and that functional role can be shared with systems that realise it in rather different internal ways.

Embedded functions can also take advantage of the work done to formulate purely attributionist theories of content.[42] Such theories formulate functional roles. To move to wide functions, all that is added is realism about mental representation, that is, a

---

40 Following the Ramsey-Lewis approach, Lewis (1970).

41 Some might hold that there is a functional role characteristic of beliefs, schematised by the idea of a representation being in the 'belief box'. Depending upon how this is spelt-out, such a functional role might specify something proprietary to content.

42 Dennett (1981b), (1987), Davidson (1974), (1984).

commitment to the existence of internal realisers. As explained in chapter 5 (subsection 4.1), this entails that there are intrinsic similarities between different tokens that realise the same representation in a given individual, and that those intrinsic similarities are relevant to the way the representation is processed within the system. So, the theorist of embedded functions can be committed to more than a purely functionalist description of contentful states. She can hold that, within an individual, the tokens which realise the same content on different occasions must be physically similar.

In short, various extant theories of content suggest one type of answer to the question of what a contentful explanation is for. They support the view that content ascription characterises the operation of a system in wide functionalist terms, allowing it to be interpreted as embedded in its environment, and giving rise to generalisations that apply across different realisations of the same embedded functions in the same environment. As such, contentful explanations can lock onto real patterns that would be invisible without taking the functionalist perspective.[43] Informational theories look only to reactions to inputs in specifying functions; wide conceptual role theories advert to both inputs and outputs. None of this shows anything distinctive about that mode of functionalist explanation which is contentful. But there may be no such litmus test. The only hint of a distinctive kind of functional role comes with causal theories, which limit the causal roles of interest to those with which the system has already been involved. Causal theories suggest that contentful ascription assigns some privileged status to the things a system has actually interacted with. That idea will recur in the following two subsections.

8.2 Conditions for Successful Operation of a Consumer Mechanism

Another approach holds that contentful explanations are only appropriate when the system to be explained has a certain special type of internal organisation, with two co-operating subsystems.[44] Representations are causal intermediaries between these two subsystems, one producing a range of representations, and the other consuming them. The action performed by the consumer system varies depending upon which representation it is presented with. As with embedded functions,[45] different representations with the same

---

43 Dennett (1991).

44 Millikan (1984), Wheeler & Clark (1999).

45 Foregoing subsection.

content in a given system must be physically similar, since the consuming system must respond to them consistently. The system of producer, representation and consumer is viewed as embedded in the environment, so that its functions are characterised in extended terms. The requirement that there are discrete producer and consumer systems puts an additional constraint on the embedded function model. However, it calls for a characterisation of what it is to be a representation producer and a representation consumer.

The idea is that a consumer system acts on the basis of the content of the intermediate representations. That is, in some way the actions prompted by a representation with a given content are appropriate to that content obtaining. But how can the actions of some system be appropriate to anything? From a naturalistic perspective, an action system just does something, which may have certain effects in the world. We could take the content of the representation to be those actual results – the results that are achieved by actions of the type initiated by that representation. But this would produce an output verificationism, parallel to the input verificationism faced by informational theories, as discussed above. It allows no distinction between true and false representations: between those that lead to results within the representation's content, and those that lead to results outside it. So, we need some other way of understanding what it is for a consumer system to act on representations. The ways to make sense of this idea that have been offered in the literature rely on there being conditions for the success of the actions carried out by the consumer system.[46] Thus, success semantics offers a substantive specification of what it is to be a consumer system.[47] There are conditions for the success of various operations of the consumer system, and those conditions vary systematically with the representations that prompt those operations. The content of a representation is then the condition for the success of the range of operations it causes. When a consumer system's output is an action in the external environment, contents will therefore be features of the external environment, or facts about it.

A theorist in a naturalistic frame of mind cannot stop there, since we have replaced one murky notion, intentionality, with the equally mysterious idea that there are, in the natural world, conditions for the successful operation of systems. Surely, that only arises for systems that have been designed; and we are after a theory of content that

---

[46] Millikan (1984), (1989), Papineau (1987), (1993), Price (2001).

[47] Braithwaite (1933), Whyte (1990), Godfrey-Smith (1994).

extends beyond artefacts. Here's where teleologists get their grip. They rely on the existence of design in the natural world, not just in man-made artefacts. Design arises from evolution by natural selection. A system operates successfully when it achieves the results it was designed to achieve. When a representation prompts a consumer system to perform a certain action, evolutionary design has in mind certain distal results. That, then, is what the representation stands for.

More carefully, the theory goes as follows. Consider the producer-representation-consumer system in the evolutionary past. The output actions performed depend upon which representation is tokened as an intermediary. Each output action could have beneficial effects on the survival and reproduction of the organism that performed it. If that organism reproduces, then the whole producer-representation-consumer system will be copied into the next generation. Thus, a causal explanation of the presence of one such representational system in the present will advert to the fact that earlier generations of that system produced beneficial effects for the organism. That gives it an evolutionary function: things that it did in the past that lead to the copying of that system down the generations and into the present. But the outputs of the consumer system only sometimes led to beneficial effects – the environment had to cooperate in appropriate ways. Consider a particular action of the consumer system which is prompted by one of the representations. For that action to have had beneficial effects, the environment must have satisfied a certain condition. That condition is what had to have been the case for the representational system to perform its evolutionary function on that occasion. According to teleosemantics, that condition is then the content of the representation: the evolutionary condition for the performance of the evolutionary function of that action of the consumer mechanism which is prompted by representations of that type. The presence of these conditions in the environment, in the evolutionary past, on occasions when representations of that type were tokened, partly explains the continued existence of the representational system today.

According to Papineau (1993), more is needed for teleosemantics to deliver determinate contents. His theory differs from that explained above. The representational system must contain two different types of representational states, corresponding to beliefs and desires. The content of desires is given by the distal results at which they are aimed, understood in evolutionary terms. For example, if the evolutionary function of a particular desire is to get food for an organism, then its satisfaction condition is the ingestion of food. Desires don't generate actions directly, but only in concert with beliefs.

The desire for food will prompt different actions depending upon the agent's beliefs: what he thinks about the situation he is in, and his instrumental beliefs about how he can get the results he is after. The content of beliefs is determined derivatively from the satisfaction conditions of desires. A given belief will cooperate with a range of desires to prompt a variety of different actions. Its content is the distal condition, common across that range of actions, that must obtain if those actions are to be successful. In a system with beliefs and desires, content is ascribed to desires by evolutionary success semantics, and is then used to derive contents for associated beliefs.

This explanation provokes a worry: why start with desires? After all, the system is symmetrical enough that the derivation could go in the opposite direction. If the content of beliefs were specified, then a parallel theory could argue that the content of a desire is just that distal result which is common to all the different actions prompted by the combination of that desire with different beliefs, in circumstances when those beliefs are true. The answer may be simply that starting at the output end allows us to naturalise content, whereas starting at the input end fails. However, the suspicion remains that a belief-desire system is more symmetrical than assumed by Papineau (1993) – that there should be something about the content of beliefs that reflects the circumstances in which they are produced, just as the content of desires reflects the circumstances in which they are satisfied.

Even without differentiating between beliefs and desires, the original model of producer-representation-consumer has a similar symmetry. But nothing has been said about what it is to be a producer of a representation. The content of a representation was fixed entirely by how it caused the consumer to act, and what the success conditions of those actions are. These theories have in common a type of answer to Field's question: what is content doing? They say the content of a representation tells you about the distal conditions for success of the actions caused by that representation. That really does add something to embedded functions, because it goes beyond what the system does, actually or counterfactually. It specifies how the world must be if those actions are to be beneficial to the system. That tells us what the system would achieve by its behaviour if the content of its representations were true.[48] The approach only considers outputs prompted by a representation, not the inputs that give rise to it. An output-oriented ascription of content will tell you what distal results a system is designed to achieve when

[48] Or, in the Papineau case, if the content of its beliefs were true.

acting on a representation, but it won't deliver predictions about the results the system will actually achieve. The content ascription specifies what the system is taking to be the case when it performs actions under the control of certain representation types. Which is to say that it specifies what results would be achieved were the things the system takes to be the case to obtain. In short, content specifies the way a truth assumption would produce predictions about what the system will achieve by its behaviour. What the theory does not tell us, if only output considerations are employed, is why we should make a truth assumption about that system at all.

If contentful ascriptions are just telling us how a truth assumption about a system would take us to predictions about the distal results of that system's behaviour, then they are of no practical interest. Indeed, it is hard to see how they can be of even theoretical interest, if they tell us nothing about how the system will actually behave, and the results it will actually achieve. The practical and theoretical motivation arises only when we are entitled to make some kind of truth assumption. That is, when the representation producer is sensitive in some way to the circumstances it is in, so that the representations it produces are true often enough. How often is enough? Often enough that useful predictions can be made about the distal results that will be achieved by the system's behaviour. That is the kind of factor pointed to by informational theories, namely, that the content of a representation should depend upon the circumstances in which it is produced. My claim is that some such input considerations should be combined with the output considerations arising from naturalised success semantics.

Millikan's (1984) teleosemantics has a role for both input and output considerations. Consumer systems are designed by evolution to produce actions that vary depending upon an intermediate representation, where each action has different evolutionary conditions for its success. But producer systems also an evolutionary function: to ensure that those intermediate representations are produced only when the condition presupposed by the consumer system's behaviour actually obtains.[49] Of course,

[49] Millikan's constraint is that when the representation is produced, the condition should obtain; not that the representation should always be produced when the condition does obtain – against error, not ignorance. But perhaps the producer system would be better viewed as designed according to both constraints. The only past occasions that contributed to survival and reproduction were when the representation was tokened *and* the condition obtained. If tokening the representation in the absence of the condition is a failing, why then is it not a failing to encounter a condition for potential successful action, and to fail to act (because of failing to token the appropriate representation)? The two are arguably on a par.

evolution may achieve this aim by a very imperfect method, often producing the representation when the condition does not obtain.[50] The representation need only coincide with the condition it represents often enough to have been useful in evolutionary history. However, the requirement that there be a system designed to produce representations is a substantial additional constraint. It shows the past behaviour of the system as not only having evolutionary conditions for success, but as having been designed to be sensitive to those conditions, so that when it does achieve a successful distal result, the mechanism by which it achieves that result will depend in part on inputs to the system having registered something relevant about its environment.[51]

Care is needed here, because the evolutionary function of the producer system must also depend, ultimately, on the entire representational system achieving beneficial results. But the idea of a representational system is one where distal results are achieved by a mechanism that has cooperating components. One component – the producer – is sensitive to inputs. When the entire system has worked successfully in the past, the producer system has reacted to inputs that carry information about the obtaining of some external condition. The producer has thereby produced a representation that is interpreted by the consumer system as indicating that that condition obtains. Similarly, when the system worked successfully, the consumer system prompted a behaviour whose success condition obtained, and did so not at random, but under the control of an intermediate representation. Granted, when we are considering functions of evolved systems in general, an action could achieve beneficial results, irrespective of what caused that action to be undertaken. However, we are here considering a special kind of evolved system – a representational system – one that has been designed to be sensitive to the conditions for success of its actions. 'Designed' to be sensitive means that a full explanation of what happened on the occasions of past success will have the obtaining of an external condition as part of the cause of the system performing that action on that occasion, as well as it being part of the explanation why that action was beneficial to the system on that occasion. That is, Millikan's claim is that it is a special feature of these

---

[50] There can be nothing good about false positives: they do not bring about beneficial effects. But nor is there anything good about false negatives: no beneficial effect follows from not representing a fact when it does obtain. Again, there is arguably a symmetry between the goals of avoiding error and avoiding ignorance.

[51] Of course, it could achieve a successful result at random, if the producer system malfunctions for some reason. But equally, the representation could be produced when its success condition does obtain, but fail to prompt a successful action because of some malfunction of the consumer system.

systems that a causal explanation of the past successful operation of the system will mention both the obtaining of some external condition, and the sensitivity of the system to that condition. An explanation which only depended on the former would underwrite an evolutionary condition for the performance of some evolutionary function, but it would not be a representational function.

The requirement of a producer system, designed to be sensitive to inputs, answers an objection to the teleosemantic project. I started with the idea of success conditions as something beyond embedded functions, that it could be the aim of contentful explanation to explain. Evolutionary functions were brought in to naturalise success conditions. But evolutionary conditions for the successful performance of an evolutionary function are not distinctive of the contentful realm. All evolutionary functions have such conditions. And, whenever performance of an evolutionary function depends upon external actions, there will be distal conditions for successful performance of that function. Where is the distinctively contentful mode of explanation? The answer, of course, is that I did not start with success conditions, but with the prior idea of a producer system, a consumer system and intermediate representations. Naturalised success conditions came in to characterise what it is to be a consumer system. But, as I argued above, we also need to characterise what it is to be a producer system. In which case, there is more to being a representational system than having evolutionary conditions for the successful performance of evolutionary functions. We also need, in explaining how the producer system has functioned in the evolutionary past, to advert to its causal sensitivity to those conditions.

There is an interesting parallel with the discussion in the previous subsection. In discussing informational theories, I argued that there must be more to content than a system being causally sensitive to some condition in the environment. That is the converse of the challenge just considered. The answer from teleosemantics is that, in addition to causal sensitivities, there must be evolutionary conditions for the success of output actions, and the two must be connected, in the content of a representation which is mediate between producer system and consumer system.

Price (2001, pp. 89-103) also argues that input sensitivities are important to a teleosemantic theory of content. She proceeds by eliciting intuitions about an example. She asks us to consider a system that produces representations at random, but then acts

upon them in systematic ways through a consumer system.[52] Even if the producer system delivered representations entirely at random, the representation could still have the function of ensuring that the output of the consumer system coincides with some condition in the environment. It can be ascribed that function in virtue of the results that are achieved when, by chance, the actions prompted are successful. Price invites the intuition that such randomly produced intermediate states are not representational. The argument above explains the source of that intuition. It derives from the fact that content attribution must also depend upon how a system reacts (or was designed to react) to inputs.

Notice that the input orientation makes content ascription useful in making predictions. The producer system's function requires it to be sensitive to some environmental condition. Provided the current environment is sufficiently similar to the evolutionary one, there will thus be occasions in the current environment when the producer system is sensitive to the fact that that condition obtains. On those occasions, a truth assumption about the representation is justified. And that truth assumption will take us to a justified prediction of the distal results of the organism's action, caused by that representation. Evolution does not ensure that the producer system is particularly reliable. It need only be reliable enough to confer a selectional advantage. However, in order to have evolved, the producer system must have been sensitive to an output-appropriate environmental condition in some range of circumstances. There must be circumstances in which the producer did detect the relevant condition, because only successes can contribute to an evolutionary explanation of the presence of the producer mechanism in the present. If the current environment is sufficiently similar to the historical one, then those circumstances will arise now; so there will be circumstances in which the producer is reliable. Thus, the need for a producer system ensures that the system realises some function from environmental inputs to distal results. Interestingly, although the motivation for teleosemantics was to supplement the idea of embedded functions, teleosemantics explains why representational systems do realise embedded functions in an important range of circumstances. They realise some ecologically-relevant embedded function because evolution has designed their internal mechanism to be a realisation of that functionalist specification.

---

[52] Price (2001), pp. 93-94.

To summarise this subsection, we have seen that teleosemantics offers a distinctive answer to the question of why to attribute content at all. Content attribution, according to teleosemantics, does more than to describe the mechanism of operation of a system, or to re-label its internal states in tractable terms, or to provide an embedded functionalist characterisation of it. Representations cause behaviour, and the content of a representation tells us the distal evolutionary condition for the success of the behaviour, performance of which is designed to be sensitive to the presence of that condition.

8.3 Projection to New Instances

A third motivation for ascribing content emerges from chapter 2. The rationale for ascribing content to clusters in a hidden layer was that only by doing so could it be explained how a network manages to project its correct performance to new samples, whose input encodings are unlike those in the training set.[53] Hidden layer clusters on their own describe something about the system's internal mechanism of operation. They show how patterns of input activity are transformed into intermediate clusters, on the way to output clusters. It is not obvious that this system need be described in contentful terms.[54] But there is another explanandum. The system has been trained in its responses to a set of patterns of input activity. It then encounters a range of novel patterns of activity, but manages to respond correctly to those inputs.[55] How does it do that? That is the question to which content ascription forms part of the answer. This is what I mean by 'projection' – not that the correct behaviour projects across a range of instances, but that there is a certain relation between the set of samples on which the network was trained and the wider set on which it can achieve correct performance. This is a very specific sense of 'project'. A system's behaviour 'projects' when there is a systematic connection between the behaviour it produced (or was designed to produce) during development, and its subsequent behaviour.

Properties of the samples are the causal source of the development in the system of a structure which implements an embedded function from those properties to

---

53 Ch. 2, ss. 3.1, 3.4 & 3.5.

54 See section (13) in part III below.

55 As judged by whether output responses elicited are appropriate to properties of the new samples that are coded into those novel inputs.

appropriate distal results.[56] If the internal mechanism of the resultant system were only characterised intrinsically, it would be mysterious how the system manages to project its correct behaviour to new inputs. However, when the system is characterised as performing an embedded function, the perceptually-different inputs are not new from the embedded perspective – they share properties with samples in the training set. Thus, properties of the training samples are the causal source of the embedded function, and that embedded function may, as a result, explain how the system responds correctly to new samples. Furthermore, knowledge of the samples and task on which the system was trained allows someone interpreting the system to predict the range of samples on which the system will perform correctly in the future.

Thus, projection is the explanandum. Properties of the training samples play three roles simultaneously: as the causal source of representational development, as the basis for predicting how the system's behaviour will project, and to explain why the system's behaviour projects as it does (why the same representations are applied correctly to new instances, which are perceptually different from the training samples, but share their distal properties).

Content ascription has the same rationale when applied to natural kind concepts, as discussed above (subsection 2.2 above). Experience with initial samples is the developmental cause of a new concept of some natural kind. The concept is the basis for projecting properties to new samples. This is explained by interpreting the concept as keeping track of some underlying feature of the samples, shared between the novel samples and those in the training set. That is, the motivation for ascribing content to the representations is to explain projection to new samples. Again, a property which figures in the content of a representation is the causal source of the development of that representational capacity, explains how behaviour projects beyond the circumstances in which it developed, and allows such projection to be predicted.

Causal theories of content may be getting at something similar (see subsection 8.1 above). They assign a privileged status to things in the actual causal history of the representation user. If I am right, the reason for this privileged role is that content is partly attributed precisely to explain how behaviour, which developed as a result of experiencing some actual objects, is projected to new objects of the same type.

---

[56] Since it was assumed in chapter 2 that outputs are contentful, output behaviours can be viewed as performing embedded functions in terms of the distal results they achieve, when correct.

In short, content attribution explains projection. However, that cannot be a sufficient condition, since each example discussed above presupposes that the task on which the system was trained can be specified. It may be a necessary condition on content attribution. Alternatively, it may merely be a happy consequence of content attribution, that in some cases it can also explain projection. To get a naturalistic grip on what it is for a system to develop a new representation to perform some task, we need an additional account. Teleosemantics, as discussed in the previous subsection, may provide it.

On closer examination, teleosemantics too has this projective element. Evolutionary forces lead to the development of a system that performs an embedded function which takes environmental conditions as input and produces distal results as output. That system's development was caused by the objects in its environment and their properties. Seeing the system as realising a function that involves those very properties allows us to explain why the system's behaviour projects into its current environment. It also allows us to predict the ways the system will respond to various environmental conditions, and the distal results that will likely be achieved thereby.

8.4 Conclusion: A Possible Synthesis

The foregoing paragraph suggests that we can form a consistent synthesis of the considerations canvassed in subsections 8.1, 8.2 and 8.3. Why go representational? To describe a system as performing a particular kind of embedded function, that generalises across different realisations. An embedded function is representational when it is realised by a mechanism consisting of a producer, a consumer and mediate representations. That cooperating system was caused to evolve or develop as a result of interaction with things in the world. Each different representational vehicle causes the consumer to output a different behaviour. The content of a representation gives the distal evolutionary condition for the success of the behaviour, performance of which is designed to be sensitive to the presence of that condition. That content explains how the representational system's behaviour can be projected beyond its causal source, and allows for prediction of how behaviour will project.

Perhaps this synthesis is a sufficient condition on contentful explanation – a sufficient motivation for explaining some phenomena in representational terms. I do not claim that it is necessary. As a sufficient condition, some of its aspects may be otiose. But it does serve to show that Field's (1978) challenge can be answered. It illustrates

some ways in which contentful explanation can be autonomous from a purely syntactic (or intrinsic-mechanistic) characterisation of a system.

(9) LOOK BOTH WAYS FOR REPRESENTATION

One of the interesting features of the theory of content in chapter 2 is that both inputs to and outputs from the system play a role in fixing content.[57] The discussion in section (8) above suggests that may be true more generally. I will take each subsection in turn.

If the role of content is to specify embedded functions mediated by the content-bearing internal states (subsection 8.1), then we should expect both inputs to and outputs from the system to form part of its functionalist characterisation. This fits with the picture of moderately externalist syntax argued for in chapter 5: syntax is the way of describing the internal workings of a system so as to view it as realising an input-output function between worldly entities.

Why should a system realise any embedded function? With artificial systems, the answer is that they have been designed to do so. For example, computers are not designed simply to carry out intricate internal manipulations; their *raison d'être* is to do something useful – to take a variety of inputs and produce relevant outputs. That is why they realise embedded functions, and why internal states can be ascribed contents specified in terms of those embedded functions. Why should we expect a system that has not been designed by humans to realise an embedded function? Teleosemantics has a distinctive answer: because the systems have been designed by natural selection to perform such functions. However, evolved systems realise many embedded functions that have nothing to do with representation, and do not call for content ascription. Teleosemantics explains this too – content ascription comes into play when the internal machinery realising embedded functions is structured into a producer subsystem and a consumer subsystem. Representations are causal intermediates between these two subsystems. Furthermore, in these cases content ascriptions do more than characterise how an internal state realises an embedded function (functionalist 'function'), they also give conditions for the successful performance of the evolved function (aetiological 'function') which the internal machinery has been designed to perform. However, this raises the possibility that content is fully determined by naturalised success semantics, and

[57] Ch. 2, ss. 6.3.

thus by a system's outputs. I argued above that teleosemantics must also take account of a system's inputs, in order to specify what it is to be a representation production subsystem. Therefore, teleosemantics also ascribes roles to both inputs and outputs in content determination (subsection 8.2).

In subsection 8.3, I suggested another role that content attribution might play. It can address a particular question about how a system relates to its causal history. The system has been caused to develop an internal structure. That may have occurred during evolutionary history, as successive generations of the system have interacted with the environment. Or, the internal structure may have developed as an individual system interacts with its ontogenetic environment. (Most likely, both processes will account for internal structure.) In both cases the system is caused, by interaction with things in the environment, to acquire the ability to perform appropriate input-output functions in respect of those worldly entities. Describing the resulting internal mechanism in contentful terms indicates the range of entities which that historical experience is likely to project to. Notice that this approach uses both inputs and outputs in fixing content. Contents show how a certain input-output function bridges the gap between historical inputs and newly encountered inputs (what I have called projection). In order to see the input-output function as projecting from historical cases to new ones, it must be characterised in wide, world-involving terms, since new inputs may share nothing except such distal properties with the historical ones. In short, subsection 8.3 gives another reason why both inputs and outputs should have a role in content determination.

These arguments are not demonstrative, but they are strongly suggestive. They may not convince the fiercest advocates of informational theories, who will persevere in trying to get enough purchase from input conditions alone to determine content. However, informational theorists would find a valuable additional source of constraint in output factors; and the arguments above suggest that there are good reasons of principle why such factors should also be relevant to content determination. The conclusion I draw is that philosophers should not restrict themselves either to input factors or to output factors as fixing content. Theorists of content should look both ways for representation.

(10) CAUSAL EFFICACY

The question of what contentful explanation is for is sometimes confused with the question whether contents are causally efficacious. These issues are separate but related. For example, suppose the answer to the former question entails that contents are functionally

specified, in externalist terms, using the Ramsey-Lewis scheme (Lewis 1970). Then their causal efficacy will depend upon whether wide functional properties are causally efficacious. That is a more general metaphysical issue, arising in relation to functional properties in the special sciences and social sciences too. Nor need a negative answer undermine the project of explaining what content is for. Contentful attribution could have autonomous explanatory validity without contents being causally efficacious. In my view, the right approach is to ascertain what contents are, such that they can play the explanatory role they do. Once characterised, we can ask about the causal efficacy of such things.

Above, some options are explained as to how this characterisation may go. Furthermore, we have the specific commitments of the theory in chapter 2. The purpose of this section is to outline a position on causal efficacy that is consistent with those views. I will not argue in favour of a particular view, because there is not space here to do justice to the complexity of published arguments about mental causation. My aim is just to locate a position in the broad spectrum of views on mental causation, and to show that it would apply to representational content, as characterised herein. The purpose is to forestall any feeling that causal efficacy presents a special problem for my approach.

Chapter 2 pointed to two important features of the connectionist case (subsection 6.4). First, clusters are properties of larger entities than the nodes and connections that form a network. Clustering is a property of an entire network (or, at the least, a significant portion of a network). Thus, clusters are found at what Kim (1998) calls a higher 'level' of entities than the component nodes, connections, weights and activation functions. Second, the content of a cluster is given in functionalist terms. That is, content is what Kim calls a higher 'order' property of a cluster.

On higher order properties (functionalist properties), my position is as follows. They are causally efficacious because their realisers are. That can be spelt out in two ways. One can argue that, when realised, the functional property is identical with its realiser. So the functional property is causal because, on each occasion when it enters the causal order it is identical with its realiser, and that realiser is causal. Alternatively, one might reject the identity, and conclude that higher order properties are not really causally efficacious, or are causally efficacious in a special way. According to this view, the functionalist description does not specify a causal property at all, but just some higher order unity that can exist between a series of intrinsically different realisations, that unity consisting in sharing the appropriate relational property. One can then hold that these

higher order properties are not really causally efficacious, or are causally efficacious in a special way: efficacy through having a realiser in a causal order.[58]

Alternatively, content may be a higher level property of a system, a property that can only be instantiated in macro-sized systems. For example, contents may be properties only of systems that are large enough to have an internal structure of producer, consumer and intermediate representations. Even if contents are specified functionally, there may be an intrinsic property shared by all the realisers of a particular content – if so, we should view the content as being that property. Such a reduction of the functionalist specification to a physical realising property would vindicate its causal efficacy. In any event, such higher level properties have to form part of the picture, since, even if contents are variably-realised functionalist properties, the realisers will be higher level properties of the system. Furthermore, my discussion of syntax gave reasons why different tokens of the same representation within a particular system should be intrinsically similar (chapter 5). Thus, at least within a given system, the tokens that realise a functionalist contentful characterisation must share a distinctive higher level property. There is a deep puzzle about the causal efficacy of these higher level properties.

That is the most substantial metaphysical issue. It arises for higher level properties of all kinds, throughout the natural and social sciences, and in everyday explanation. Higher level properties are found that just do not exist at lower levels. One example is flammability – that cannot be predicated of single molecules. Another example is stereo-isomerism in chemistry. Stereo-isomerism cannot arise in atoms, but only in some molecules. Where it exists, it has significant causal consequences (for example, the laevo form of some sugars is not digested). But stereo-isomerism is determined fully by the spatial arrangement of component atoms. The problem with higher level properties in general is that the properties of a complex entity seem always to be determined, metaphysically, by the properties of its components and their relations (Kim's 'mereological supervenience', 1998). But those components and their properties are also found in the causal order. So we seem to have two rival accounts of how the universe unfolds. According to one, higher level properties can be causally interrelated: M1 causes M2 one second later, say. According to the other, their components are causally

[58] A different objection argues that there are two candidates for causal efficacy even at the level of the realiser: a particular realiser, or the disjunction of the possible realisers. In my view, the latter will not generally be a property capable of causal efficacy, for the reasons given in ch. 2, ss. 3.5.

interrelated: physical laws operating on each of the components specify where those components will be one second later, what properties they will have, and how they will be interrelated. The physical configuration that is the result of the operation of the lower level causal laws[59] determines[60] that the system has property M2. So, the puzzle goes, which is causal? The causal relations between higher level properties? Or the causal interrelations between their components, on which those properties mereologically supervene?

Here is the sketch of a suggestion for an answer. Causation is not just a matter of necessary connection between properties, but something more. Causation requires properties to be related as a matter of natural law. Natural laws describe real patterns in the world, and those patterns can be found at many levels, from the micro to the macro. The same region of space-time may take part in a whole range of such patterns, and many others may be instantiated within it. Some of these patterns exist between entities at different levels. When there are synchronous strict laws between levels, then the properties at the higher level reduce to those at the lower level. However, often there are only *ceteris paribus* bridge laws between levels. In that case, moving up or down several levels may destroy any law-like connections between properties at those levels, even *ceteris paribus*. That is because the *ceteris paribus* conditions can be different for bridge laws at different levels, so that they wash out, making no *ceteris paribus* condition available across several levels. That makes the levels completely autonomous. We can hold onto mereological supervenience, but such necessary dependence of the higher level properties on properties and relations of lower level entities does not amount to a lawful connection. The picture is of real patterns at all levels, with autonomy between many of the levels, and without a privileged level at which to describe the patterns.[61] Since there

---

59 Whether those laws are deterministic or probabilistic, the later lower level arrangement of components is the causal result of their earlier lower level arrangement.

60 'Determines' here is constitutional, not causal, and is the result of mereological supervenience.

61 It might be asked whether the supervenience chain 'bottoms out' at the level of the most basic, fundamental physics. If so, does that bottom level, upon which all other properties supervene, not have a privileged status? I have two answers. First, even if it has some privileged status, that need not deny causal efficacy to other levels too, since supervenience is not reduction. Second, there are reasons to think that the bottom level, upon which all else supervenes, is not causal either. Plausibly, the most basic physics is not causal. It merely specifies how various fields can be distributed in space-time. That these field equations are not causal laws is reflected, I would argue, in the fact that they are time-reversible.

is no privileged level from which to look at the patterns, there is no privileged level of description, therefore no privileged level at which causation occurs.

Clearly, these are deep waters. To defend the current thesis, I rely only upon two claims. Firstly, if contents are functionalist properties, then they are least as causally efficacious as any other functionalist properties; and their importance as an explanatory scheme should be elucidated whether or not they are causally efficacious. Secondly, if the contents are higher level properties of representational systems (or are causally efficacious in virtue of being realised by such higher level properties), then they *are* causally efficacious, since higher level properties can be autonomously causally efficacious. To say why depends upon an answer to a deep metaphysical puzzle, turning on the nature of causation itself. There almost certainly is an answer, since the causal efficacy of higher level properties arises in so many fields. At the very least, contentful properties are in no worse position than the properties that figure in causal explanations in all of the natural and social sciences.

III. Reliance on Historical Factors

(11) WHY WON'T CURRENT FACTORS DO?

At various points in the thesis I have relied upon accounts of content that depend upon historical factors – that the content of a representation is determined, in part, by things in the past. Teleosemantic content depends upon facts about an organism's evolution (subsection 8.2 above). My theory of content for connectionist systems in chapter 2 makes content partly dependent on facts about the circumstances in which a system developed. Furthermore, I argued in part I of the current chapter that developmental circumstances may figure, as content-determining, in an appropriate theory of representation for some other systems. In this final part of the chapter, I address an objection that can be made to

---

These are deep issues in the philosophy of physics. But here is the speculative idea: that there is an underlying non-causal fabric of space-time, described by field equations. Fundamental particles are then patterns in these fields, as are all higher level entities and properties. The machinery of causation, properties and natural laws is used to capture these patterns, which can occur autonomously at various levels. Thus, there is no privileged basic level of causation.

any theory which makes content dependent on historical factors. The worry is: won't current factors do?

The worry arises because of realism about representation. The vehicles of content are thus current physical entities. Each such entity could be picked by its current properties, for example, spatio-temporally. (That the entities can be picked out one by one in current terms, for example by their space-time location, does not entail that any current property is shared by and distinctive of those entities, so as to be sufficient to individuate the class.) Furthermore, these entities must interact with the world, and with each other, in ways that are at least sometimes faithful to their contents. Surely current properties of the vehicles of content can explain those interactions? But if vehicles can be individuated by current properties, and interact in virtue of current properties, why is there not an adequate way of specifying content in terms of current properties? That is the concern.

I have already argued that intrinsic properties of a system taken in isolation do not determine its contents (chapter 5). Amongst various intrinsic descriptions of the mechanism, the one that specifies vehicles of content must fit with how the system behaves in its environment – the syntax is the mechanism which implements some embedded function. Such embedded functions have the added virtue of generalising across different systems where they are variably realised. Intrinsic properties of a particular system could not, therefore, support such generalisations. But that is an argument for moving to wide functions. It is neutral on whether the functional characterisation should depend upon historical factors, or only upon current ones.

One answer is that even specification of internal mechanism depends upon historical factors, since it necessarily proceeds by identifying historical functions. I expressed unease with that idea in chapter 5, but could not reject it definitively (subsection 14.2, chapter 5). To underpin representational realism, syntax must be causally efficacious. This counts against syntax being specified historically. That is not because historical factors are causally inefficacious – of course they have effects in the present. Rather, the worry is a strong metaphysical commitment to the causal priority of the present. Present causes screen off past causes. Where past causes have a current effect, that occurs in virtue of current causes. This metaphysical view is held widely and deeply. As is common with such fundamental claims, it is hard to find independent philosophical justification for it. It derives most strongly from the framework of natural science, which is committed to the idea that the total current physical state of the

universe determines all that there is to be determined about what will happen next. Differences in the past have no impact on the future unless they are reflected in some difference in the present. Certainly, a large body of empirical research supports the claim that the basic laws of physics, at least, are blind to history in this way. From these considerations derives my reluctance to accept that syntactic properties, which must be causally efficacious, should be dependent upon historical factors.

Accepting that the biological functions of an evolved system depend upon history does not entail that syntax should also do so. As well as describing a system in terms of its normative, teleofunctions, it seems perfectly possible to describe its mechanism in terms of Cummins functions (1984; 1996, ch. 8).[62] This is a liberal notion. A complex system can be divided up many ways into interacting entities, and can be seen as interacting with the environment in many different ways. Each will go with different specifications of Cummins functions for the system and its components. None is preferred over the others. It is a purely descriptive exercise. The challenge to historical syntax is that one of these Cummins-functional descriptions might be just as good for prediction and explanation as anything specified in historical terms.

The challenge has been posed graphically with a thought experiment. Davidson invites us to imagine that, by an incredibly unlikely coincidence, a bolt of lightning in a swamp produces a molecule-for-molecule duplicate of himself (Davidson 1987). The swampman would immediately behave just like Davidson, and its internal workings would be the same. Davidson points out that, according to his theory of content, swampman has no contentful states. Millikan makes clear that her theory of content entails the same conclusion (Millikan 1984, p. 93; 1996a). But surely swampman must have contentful states? If so, swampman is an objection to historically-based theories of content.

The thought experiment trades on the strength of our commitment to the causal priority of current properties. The intuition is that swampman would act, speak and feel just like a normal human. No historical difference between swampman and Davidson could make a difference to its future, since at the time of the lightning bolt they were intrinsic duplicates. Thus, the thought experiment relies upon the fact that current causes screen off historical ones.

The swampman thought experiment is so exotic that it can be misleading. It invites misplaced responses. I will mention some in the next section. Therefore, it is

---

[62] Millikan (2001) discusses the potential of Cummins functions to provide rival functional explanations.

usually best to focus on the underlying problem, which is the following. Even if historical properties are used to individuate contentful entities, and to classify them together on the basis of their contents, why are there not current properties of those entities (including the current-embedded functions of the systems in which they are found) which can serve the same explanatory goal just as well, or better?

The answer will depend upon what contentful ascription is for. That is, an adequate response will depend upon an answer to the question posed in part II, of why contents should figure in explanations at all. I suggested three lines of response above, with contents as: specifying embedded functions performed by a system (ss. 8.1); giving conditions for successful actions, to which the system is sensitive (ss. 8.2); or explaining the projection of behaviour into circumstances beyond those in which it developed (ss. 8.3). Subsections 8.2 and 8.3 both made use of a further idea, which gives an epistemological role to causal history. Historical factors cause the development of representations in the system. Therefore, knowledge that a system has the right sort of causal history justifies a claim that it has contentful states; and knowledge of the detailed circumstances of that development justifies the ascription of particular contents.

I did not arrive at a definitive conclusion as to the purpose of content attribution, so these suggestions are employed here as placeholders. They show how answers to the issue in part II contribute to assessing the permissibility of relying upon historical factors. However, only when it is clear why we go representational can it be definitively assessed whether current factors could be adequate to that job, or whether historical factors must be relied upon.

Section (12) discusses immediate responses to the swampman thought experiment, some of which can be dismissed. Section (13) gives an answer to a swampman-type objection to my theory of connectionist content in chapter 2. Section (14) considers some of the ways that advocates of teleological theories of content justify their reliance on historical factors. Section (15) draws out two general themes from these answers.

(12) First Responses

A common response to swampman is to claim that he is physically impossible, or too improbable to be worth considering. Another response argues that it is entirely unobjectionable that he should have no beliefs or desires since, were he to have any, they would mostly be false anyway. These quick responses may answer the thought experiment, but they miss the underlying problem. They don't tell us why historical

factors are better than current ones in determining content. In a similar vein, an early response from teleosemantic theorists was to argue that the intuitions elicited by swampman are defeasible, since the theory offers a theoretical reduction of the concept of content, not an analysis of it.[63] That objection misses the point, since the thought experiment is not relied upon as a source of intuition, but rather as a graphic illustration of the consequence of combining teleosemantics with an uncontentious commitment to the causal primacy of current properties.

Surely, it is true of any actual human being that, if she did not have an evolutionary and/or developmental history, she would not have any contentful states? Yes, but – swampman highlights the status of that counterfactual. Any human has representational capacities because she has developed in an appropriate environment, and because she has evolutionary ancestors. But that causal fact does not alone entail anything about the metaphysics of content.

Perhaps there is a simple response, that takes historical factors as sufficient, but not necessary. Historical considerations are used to individuate some current entities, and to give their contents. But anything physically similar to the current state of an evolved/developed system has the same content as would be determined historically. That is to say, contentful explanation covers the following class of systems: 'historical' systems (any system with an appropriate causal history, 'appropriate' specified by the theory of content), and any system that is physically similar to the current state of an historical system.[64] A first problem with this line is that it is not clear that the required physical similarity can be specified in non-contentful terms. Even if it can be, the challenge remains: the similarity relation specifies some current physical properties – why not just use those to individuate content? Historical factors may be a way of arriving at the relevant current physical properties, but this account does not rule out content being determined, metaphysically, by those current properties.[65]

---

[63] Papineau (1993, p. 93), Papineau (1996), Millikan (1996a), Neander (1996).

[64] Thanks to Matteo Mameli for pointing out this possibility. A similar position is spelt out by Michael Tye (1998) at p. 463; although he goes on to argue that it won't do the work he needs to allow representational content to account for the qualitative character of sensory experience.

[65] Indeed, it may follow from this line that the appropriate reaction to swampman is to accept the objection, and therefore to specify a way of individuating content in terms of current properties.

A stronger defence of historical factors would show that content is fixed in part by historical properties, showing how the metaphysical relation between historical factors and content follows from the very nature of contentful explanation. That is what is sought in the remainder of the chapter. Should that fail, however, we can fall back on the more modest role for history suggested in the last paragraph: whether a system has contentful entities, and the content of those entities, depends only upon the current physical properties of that system; but historical factors are a way of individuating some such contents, in the sub-class of contentful systems that do have an appropriate history.

(13) CONNECTIONIST SYSTEMS

Could a swamp connectionist network have contentful states? Are current factors adequate to determine the content of states of a connectionist system? The argument for attributing content to the states of a connectionist system relied upon it having a developmental history (ch. 2, ss. 3.1 & 3.4). Only by seeing the behaviour of the hidden layer in contentful terms could clustering in the hidden layer help to explain how the system managed to project correct performance to novel samples. That is, the explanatory project requires a developmental history. However, perhaps content ascription can be addressed to a weaker explanatory task: to explain how the system currently behaves, without being committed to whether that behaviour is a novel projection from the training set. Hidden layer clustering (i.e., clustering in state space of activity produced in response to some set of inputs) clearly describes a feature of the network's processing, irrespective of how the samples are selected. In particular, the hidden layer state space could be plotted on the basis of some set of samples to which the network responds correctly, irrespective of whether they formed part of the training set (indeed, this could be done without knowing what the training set consisted of). If there were clusters in this state space, then should content be ascribed to them, using the principles of 3.5.1 (chapter 2)? That kind of content ascription would not make content dependent on developmental history. Doubtless, the network would have to have had a developmental history in order to show interesting correct behaviour. But nothing about that history would constrain content ascription, if this proposal were sustainable. As before, content would have to be pre-assigned to the outputs in some way. But that alone would give a basis for ascertaining a set of samples on which the network performs correctly; and thus a basis for individuating clusters in hidden unit state space.

The motivation for ascribing content is not as strong in this case – the original motivation relied upon in chapter 2 is weakened. Of course, the clustering might be an interesting feature of the system. However, those clusters could equally well be described as clustering together patterns of input layer activation (not mentioning anything about the real-world samples). Clusters still provide a way of describing the operation of the system that abstracts away from individual patterns of activation. But, without the projection to be explained, it is unclear why an explanation need move beyond viewing clusters as an intrinsic feature of the mechanism of the system. Certain input patterns result in correct outputs. The network achieves that in part by clustering in the hidden layer. None of that explanation need advert to the external samples, or their properties.

Perhaps it would be legitimate still, in such circumstances, to re-label the clusters with input properties, chosen according to 3.5.1. If so, there would be a species of content ascription available for such systems which is not constrained by developmental history. However, the motivation for seeing such ascription as genuinely contentful is weaker.[66]

Thus, the best argument for attributing content to states of connectionist systems does not have content fixed by current factors alone. It also depends on ontogenetic factors: the particular samples and task on which the network happens to have been trained. This answer connects with the reasons for attributing content suggested in part II above (section (8)). Current properties of a connectionist network can license the ascription to it of embedded functions (ss. 8.1), and of conditions for successful action (ss. 8.2). The latter would derive from the way outputs of the current system are interpreted, irrespective of how they were interpreted during training. All that is missing is a role for content in accounting for projection (ss. 8.3). Whether current properties can determine content for connectionist systems depends upon whether the third suggestion is necessary to legitimate content attribution.[67]

Notice that history can play an epistemological role here, as foreshadowed at the end of section (11) above. If we know nothing about a network, we have no reason to think that it will display any interesting behaviour, or implement any interesting embedded function. Nor are we justified in expecting there to be any higher level features of its

---

66 Ch. 6, sec. (13).

67 Or whether some other motive for its legitimation, not considered here, necessitates an historical approach.

internal organisation, like clustering.[68] Conversely, knowing that a system has been trained on a certain sample set to perform a certain task tells us quite a lot about it, and would justify attempting to explain its behaviour in contentful terms. This epistemological connection to history does not entail that historical factors are content determining. It does reflect the fact that, in a connectionist system with a developmental history, the circumstances of that history (output task, properties of the training samples) are the causal source of the structure that exists in the current system.

In sum, the connectionist model suggests that content is partly determined historically, because part of the purpose of content attribution is to explain projection from historical training to present behaviour (as argued generally in subsection 8.3 above). It also illustrates the epistemological role played by knowledge of history in attributing contents.

(14) Answers From Teleosemantics

14.1 Wider Generalisations

Millikan defends historical categories on the basis that they underpin a wider range of generalisations than could be made on the strength of current categories (Millikan 1984, pp. 93-94; 1996a). She draws an analogy with biological species. Species figure in many biological laws. Members of a biological species must share a common ancestor, so they are individuated partly historically. This way of classifying species is found to be the most useful basis for biology.

Millikan argues that humans and swamp people do not fall together under any natural kind. Historical content ascriptions work for prediction and explanation in the historical kind *human*. There is no reason to think those generalisations will carry across to swamp people, who materialise by chance in an instant in a primordial swamp.

That certainly makes an epistemological point. We have no reason to think that a swampman, at the moment of creation, has contentful states. Whereas, when we encounter another person, we can infer from her membership of the natural kind *human* that she has contentful states, because members of the historical kind *human* share a history, which we know includes the evolution of representational systems. As Millikan argues: the inference from *human* to *content-bearing* is empirical, whereas an inference

[68] See ch. 2, ss. 6.1.

from *swampman* to *content-bearing* can at best be logical, deduced from swampman being an intrinsic duplicate of a human (that fact is designed into the thought experiment). So the idea must be that epistemological considerations constrain the metaphysical nature of content, and require it to be historically determined.

It is also clear that, in the actual world, a biology of a category that covered humans and their history-free intrinsic duplicates would be no more useful than current biology, since there are no history-free intrinsic duplicates of humans in this universe, since they are mind-bendingly unlikely to arise. But that is a false opposition, arising from taking on the swampman thought experiment too directly. The choice is between an historical-functional characterisation of content and one which characterised embedded functions in purely current terms. Without a history, swampman could instantiate embedded current functions. They can also ground the epistemology: after watching swampman for a while, talking and acting appropriately in the world, we would have reason to think that he does instantiate current embedded functions, irrespective of his history. The challenge of swampman is not to formulate categories that can cover actual organisms without a history. There are none of those. The challenge is that embedded functions might provide generalisations that are just as good, or better than those underwritten by historical functions, and those generalisations would be true irrespective of the history of the systems to which they apply. If so, current and historical functions are not on a par. Because, recall the deep metaphysical commitment behind the thought experiment. It is common ground between those in favour and against historical factors that swampman will act and react just as his human double would do, in the same circumstances. That is, it is agreed on all sides that current causal factors screen off historical ones. That is a good reason to accord metaphysical priority to the current properties, when we are asking on what basis, metaphysically, content is determined.

In short, it is far from clear that historically-based generalisations are more well-founded metaphysically than ones based in a system's current properties. However, knowledge that a system has a human evolutionary history does allow one to infer that it has representational systems, and thus to infer things about its behaviour. More specific knowledge about circumstances of evolution and development can ground more detailed inferences. Millikan's defence of historical factors seems to amount to the claim that such epistemological factors require content to be historical. That is a perfectly tenable position. It goes with a particular view of what content attribution is doing. Recall that one answer to Field's (1978) puzzle is that contents provide a convenient re-labelling of

the syntax of a system (see subsection 7.2 above). That re-labelling helps us humans to understand how a system operates, since it re-describes things in terms that we are used to, and are useful for our purposes. If content is convenient re-labelling, then it is easy to see why epistemological considerations should affect which re-labelling we should use. But the theorist who pursues that line must accept this relatively modest characterisation of the nature of content attribution.

There is an alternative way of attacking the idea of content as specified by current embedded functions – there are just too many. Current embedded functions are not sufficient because swampman would realise very many different sets of current embedded functions, if he were located in different environments. Think of him like a computer, designed to play chess, but used to implement military strategy. Provided the function from proximal inputs to proximal outputs is appropriate, those inputs and outputs can be connected to a variety of distal situations. Of course, in practice we are only interested in the behaviour of swampman in the same sort of environment as that occupied by the molecular duplicate on which he is based. In that particular environment, swampman will not be realising any exotic embedded functions. But why chose that environment? After all, swampman has no history, so there is nothing about him that connects him to any particular actual or counterfactual environment. What gives that environment its special status is that it was the environment in which swampman's molecular duplicate evolved and developed so that he realises embedded functions. Thus, even in the swampman thought experiment, history is playing an epistemic role: it is telling us which environment we should embed swampman in, if he is to realise what are intuitively the right embedded functions.

14.2 Naturalising Intentionality

Papineau (2001) defends teleosemantics in a slightly different way. He says that folk psychology characterises contentful states. It specifies various 'roles' that they play in psychological prediction and explanation. The job of a theory of content is to explain how entities in the world, characterised naturalistically, can fulfil those roles. Teleosemantics provides such a naturalistic, reductive theory, applicable to our world. That is to say, the folk psychological content 'role' is filled, in our world, by historical properties – those characterised by teleosemantics. 'Role' and 'realise' cannot be meant here in their usual functionalist sense, where to realise means to move from entities characterised relationally (functionally), to an intrinsic characterisation of one entity that has the

specified relational property. Instead, I think Papineau uses 'realising the folk psychological role' as an intuitive shorthand for 'providing a reduction, appropriate in the actual world, of the phenomenon which is roughly characterised by folk psychology'.

Neander (1996) makes an argument that is superficially similar. She justifies teleosemantics as the best alternative for capturing the phenomena of intentionality: for ascribing non-disjunctive, normative truth conditions, which agree with commonsense about representational content in most cases. That presupposes a rather shallow answer to the question of what a theory of content is for. The purpose of the theory is just to show how the phenomena of intentionality can result from the non-intentional properties of the natural sciences. No light is thrown on the question of why we should attribute content in the first place.

Papineau's argument is not so quick. He explicitly disclaims the need to show that historically-based contents are better at solving disjunction problems, or accounting for normativity, than theories based on current properties.[69] His project does not require there to be anything metaphysically preferable about historically-ascribed contents. Historical factors are part of the appropriate reduction in our world. That is because the systems which in our world behave in the ways described by folk psychology, do so because of their evolutionary history. And there are good reasons for that. As a matter of physical possibility, it is very unlikely that such systems could arise any other way.

I argued above (section (11)) that an adequate answer to the challenge to historical factors should rely upon an appropriate answer to the question of what contentful explanation is for. The type of response given here has not, so far, made that connection. But it can be made. Recall, that Papineau's teleosemantics has an answer to the question (subsection 8.2): contents tell us about the success conditions for a system's actions. In the actual world, the only systems that have success conditions have evolved (or have been designed by humans). So the appropriate reduction of success conditions proceeds historically. Because that is what content is up to, contents are historically determined. If we were not dealing with systems designed by evolution, there would be no question to which content is the answer.

---

[69] Papineau (2001, p. 280).

14.3 Projection

A third style of answer is also possible. Suppose contents are attributed, in part, to explain how a system manages to project its behaviour from samples that caused its representations to develop (or caused its representational systems to arise in evolution), to current circumstances (subsection 8.3). Then, in a system without a history, there is no such explanandum. If projection is necessary to justify going representational, then systems without a history are simply not susceptible to contentful explanation. We may treat them as if they have contents. That may perhaps provide a convenient re-labelling of their syntactic states. But the reason for moving to a contentful level of explanation would be absent.

(15) TYPES OF ANSWER: WHY RELY ON HISTORICAL FACTORS

The previous section gave a partial survey of answers to the objection to historical factors as content determinative. The three styles of answer (14.1, 14.2 & 14.3) align roughly with the three motivations for content attribution discussed in part II above (subsections 8.1, 8.2 & 8.3, respectively). These answers can also be classified along another dimension.

Some responses rely upon broadly epistemic considerations. Knowledge about a system's history allows us to infer things about its contentful states, and thus about how it will behave. Historical factors allow us to make more useful generalisations. Perhaps I can capture these considerations as follows: a system's history gives an epistemic basis for thinking about it in a certain way. And that epistemic relation obtains because the history is the causal source of the structure which is described in these historical terms. It has not been fully explained why these historical characterisations are any better than embedded current functions. But perhaps that can be answered by reference to the generalisations that are useful or appropriate in the actual world (where representational systems have arisen through evolutionary design). If contents are merely an appropriate re-labelling of entities for the purpose of this epistemic project (the project of inferring from history, or making generalisations), then the objection to historical factors loses its force. Current causes may screen-off historical ones, but there is no reason to think that convenient current labellings should displace convenient historical labellings.

Other responses see contents as answering a particular type of question, where the question fails to arise if a system has no appropriate history. Thus, if contents are to

characterise success conditions, there are no success conditions for the behaviour of a system unless it has an evolutionary history. Or, if contents are to explain projection from historical samples (samples which were the cause of representational development or evolution) to current circumstances, then there is no explanandum if the system has no history. These answers share the view that, in respect of systems without an appropriate history, there is no explanandum to which contentful explanation is the explanans. Non-historical systems can still have current embedded functions. But embedded functions are not distinctive of the contentful realm. The idea is that we need to be explaining historical systems before what is distinctive about contentful attribution can get its grip. The primacy of current causes is irrelevant, since in a system characterised in terms only of its current properties, there is no reason to go representational.

Perhaps these two considerations can be combined. Maybe content attribution arises just in those cases where a system's history raises certain explanatory questions (success conditions, projection), and where the epistemic basis for an answer to those questions is the existence of that history.

(16) CONCLUSION

This part of the chapter attempts to answer a swampman-type objection to the reliance, elsewhere in the thesis, on developmental or evolutionary circumstances as partly determinative of content. I have argued that the objection cannot be avoided just by disposing of the swampman thought experiment. A good answer should give a positive reason for relying upon historical factors. A full answer will be connected to an idea of what content attribution is for.

I have not advocated a definitive answer. Instead, I have shown how an appropriate answer can be formulated in the light of different views about the fundamental nature of content attribution. However, the various considerations in favour of historical factors do, collectively, amount to a strong argument against swampman being a definitive objection. In particular, if content concerns success conditions (subsection 8.2) or projection (subsection 8.3), then it is clear why content should be partly determined by historical factors.

In order to support the claims that I have made in this thesis, I need only the more modest conclusion: not that content is determined by historical factors, but that it may be. That is enough to rebut the claim that swampman is an insuperable objection to my theory of connectionist content (chapters 2 and 3), or to my arguments in part I of chapter

6 that the content of representations in some systems may be partly determined by the circumstances in which they developed. Given our current imperfect understanding of why we go to a representational mode of explanation at all, it is at least an open possibility that content should be determined in part by historical factors. In fact, in the light of some reasonable accounts of the purpose of contentful explanation (part II above), there are positive reasons for thinking that such explanations can only be given of systems with an appropriate history. If so, there is no objection to historical factors being metaphysically necessary for a system to have contentful states.

7

Conclusion

To conclude, I briefly summarise the progress that has been made in the course of the thesis and suggest some avenues for further research.

The theory of content for connectionist systems in chapter 2 derives strong empirical support from existing work in computer modelling and is, I argue, theoretically compelling. However, it will only ultimately be vindicated if it proves to be applicable and useful across a range of connectionist models. That is a topic for further research, as are the empirical predictions made in that chapter. The research would best be carried out through collaboration between philosophers and cognitive scientists.

Chapter 3 suggests how my theoretical approach can be extended to other connectionist networks. There is more theoretical work to be done here, to formulate detailed theories for each of a number of actual connectionist models; as well as empirical work to test the applicability of the theory in these cases. Much more could also be said about appropriate theories of content for simple representational systems in real biological brains. It is realistic to hope that the approach advocated in chapter 2 and 3 could lead to detailed philosophical theories of content for some such simple systems, where the systems are already well-described in the empirical literature. Chapter 3 contains a strong enough case for my approach to motivate detailed philosophical work in this area.

The points about prototype theories of content in chapter 4 constitute a settled philosophical view. However, the practical importance of connectionist models of typicality effects will become clearer as empirical work continues. I explain typicality effects in connectionist systems by relying upon a description of syntax in terms of clusters. That argument will only be compelling when it is shown, across a wide range of models, that typicality effects do arise because of the structural features I describe.

Chapter 5 suggests that, in thinking about content in general, theorists should not take syntax for granted. In fact, as the case study in chapter 2 shows, getting clearer about a system's syntax can inspire a better approach to content in that system.

Chapter 6 sets a philosophical challenge that merits much further investigation. Consistency with the mechanisms of representational development is a substantial constraint on theories of content. Developmental circumstances may even partly determine content. I have given a detailed argument about how that should work in the context of teleosemantics. Further research should explore how developmental circumstances can be taken into account by other approaches to content. Ultimately, by pursuing this line, we may arrive at a structure for a theory of content for all kinds of mental representation. This is the chapter of the thesis that generates the most avenues for further philosophical research. The research programme that it suggests is very substantial indeed but, I would argue, those difficulties are in proportion to its importance.

Chapter 2 pointed out three interesting features of my theory of connectionist content. I have mentioned the moderately externalist syntax and the content-determining role that it suggests for development. The third feature is less unorthodox: the theory gives a role to both inputs to the system, and its outputs, in fixing the content of its internal states. Other theories do the same. For example, wide conceptual role semantics has content functionally specified, usually in terms of both inputs and outputs. Chapter 6 suggests that might be a constraint on any adequate theory of content (sections (8) and (9)). I do not make a general argument to that effect. However, the three motivations for attributing content that I discuss (subsections 8.1, 8.2 & 8.3) all suggest that both inputs to and outputs from a system are content determining. That conclusion is uncontentious in two cases: content as specifying embedded functions, and content as explaining the projection of a pattern of output to new input samples. However, it is less clear that teleosemantics should allow that content partly depends upon how a system is sensitive to the environment it is in. Different theorists have different views. Subsection 8.2 of

chapter 6 argues that teleosemantics should make content partly dependent upon inputs. These three examples amount collectively to a strong reason for thinking that theories of content should advert to both inputs and outputs as content determining. That is, theorists should look both ways for representation.

Part III of chapter 6 rebuts swampman-type objections to reliance on historical factors. However, a conclusive answer to the objection to historical factors must depend upon a settled answer to the question of why to go representational at all (considered in part II of chapter 6). That issue clearly merits further philosophical research, which should be pursued in conjunction with attempts to formulate improved detailed theories of content.

ACKNOWLEDGEMENTS

The detailed feedback of my supervisor, David Papineau, has been invaluable in developing this thesis. I am very grateful for his inspiration, and for his generosity with time and criticism. I am also grateful for helpful comments on some of the above material from Andy Clark, Peter Goldie, Matteo Mameli, Sarah Patterson, Jesse Prinz, Richard Samuels, Gabriel Segal and Michael Wheeler; and audiences at King's College, London, and the 2002 E.S.R.C. Workshop on Categorisation, Recognition and Perception.

The author gratefully acknowledges the support of the Arts and Humanities Research Board for the research reported in this paper.

BODY TEXT: 76,000 words

TOTAL TEXT: 86,750 words (including footnotes, references, headings, abstract and contents)

DATE: 20 June 2003

References

Ariew, A. (1999). 'Innateness is Canalization: In Defense of a Developmental Account of Innateness'. *Where Biology Meets Psychology*. V. Hardcastle. (ed.) (Cambridge, MA, MIT Press).

Armstrong, D. M. (1983). *What is a Law of Nature?* (Cambridge, CUP).

Armstrong, S., L. Gleitman, et al (1983). 'What Some Concepts Might Not Be.' *Cognition* 13: 263-308.

Atran, S. (1990). *Cognitive Foundations of Natural History*. (Cambridge, CUP).

Atran, S. (1999). 'Itzaj Maya Folkbiological Taxonomy: Cognitive Universals and Cultural Particulars'. *Folkbiology*. D. Medin and S. Atran. (ed.) (Cambridge, MA, MIT Press).

Atran, S. (in preparation) [Amongst Itzaj Maya people, genetic or kinship properties do not appear to be definitive of membership of biological categories.].

Barsalou, L. (1999). 'Perceptual symbol systems.' *Behavioural and Brain Sciences* 22: 577-660.

Barsalou, L. (2003). 'Situated Simulation in the Human Conceptual System.' *Language and Cognitive Processes* Special Issue on Semantic and Conceptual Representation.

Bateson, P. P. G. (1966). 'The characteristics and context of imprinting.' *Biological Review* 41: 177-220.

Bechtel, W. and A. Abrahamsen (2002), *Connectionism and the Mind*, 2nd ed. Oxford: Blackwell.

Berkeley, I. (2000), 'What the #$*%! is a subsymbol?', *Minds and Machines* 10, pp. 1-14.

Berlin, B. (1978), 'Ethnobiological Classification' in *Cognition and Categorisation,* E. Rosch and B. Lloyd (eds.), (Hillsdale, NJ, Laurence Erlbaum).

Block, N. (1986). 'Advertisement for a Semantics for Psychology'. *Midwest Studies in Philosophy, vol. 10: Studies in the Philosophy of Mind*. P. A. French, T. Uehling and H. Wettstein. (ed.) (Minneapolis, University of Minnesota Press): 615-678.

Bontley, T. (1998), 'Individualism and the Nature of Syntactic States', *British Journal for the Philosophy of Science* 49, pp. 557-574.

Braithwaite, R. B. (1933). 'The Nature of Believing.' *Proceedings of the Aristotelian Society* 33: 129-146.

Brooks, R. (1991), 'Intelligence without Reason', in *Proceedings of the Twelfth International Joint Conference on Artificial Intelligence*, pp. 569-595. San Mateo, CA: Morgan Kauffman.

Burge, T. (1986), 'Individualism and Psychology', *Philosophical Review* 95, pp. 3-45.

Burgess, N. and J. O'Keefe (1996), 'Neuronal computations underlying the firing of place cells and their role in navigation', *Hippocampus* 6, pp. 749-762

Calvo Garzón, F. (2003). 'Connectionist Semantics and the Collateral Information Challenge.' *Mind & Language* 18(1): 77-94.

Carey, S. (1985), *Conceptual Change in Childhood*. Cambridge, MA., MIT Press.

Carey, S. and F. Xu (2001). 'Infants' knowledge of objects: beyond object files and object tracking.' *Cognition* 80: 179-213.

Chomsky, N. (1959). 'Review of Skinner's *Verbal Behavior*.' *Language* 35: 26-58.

Christiansen, M. and N. Chater (1992), 'Connectionism, learning and meaning', *Connection*

Science 2, pp. 53-62.

Churchland, P. M. (1981), 'Eliminative Materialism and the Propositional Attitudes', *Journal of Philosophy* 78, pp. 67-90.

Churchland, P. M. (1991), 'Some reductive strategies in cognitive neurobiology', in *A Neurocomputational Perspective: The nature of mind and the structure of science*, pp. 77-110. Cambridge, Mass: MIT Press.

Churchland, P. M. (1993), 'Fodor and Lepore: State-Space Semantics and Meaning Holism', *Philosophy and Phenomenological Research* 53, pp. 667-672.

Churchland, P. M. (1996), 'Second Reply to Fodor and Lepore', in R. McCauley (ed.), *The Churchlands and Their Critics,* pp. 278-283. Cambridge, Mass: Blackwell.

Churchland, P. M. (1998), 'Conceptual Similarity Across Sensory and Neural Diversity: The Fodor/Lepore Challenge Answered', *Journal of Philosophy* 95(1), 5-32.

Churchland, P. M. and P. S. Churchland (1983). 'Stalking the Wild Epistemic Engine.' *Nous* 17: 5-18.

Churchland, P. S. and P. M. Churchland (2002), 'Neural worlds and real worlds', *Nature Reviews Neuroscience* 3, pp. 903-907.

Churchland, P. S. and T. J. Sejnowski (1989), 'Neural Representation and Neural Computation', from Nadel, Cooper, Culicover & Harnish (eds.), *Neural Connections, Mental Computations*. Cambridge, Mass: MIT Press.

Churchland, P. S. and T. J. Sejnowski (1992), *The Computational Brain*. Cambridge, Mass: MIT Press.

Clark, A. (1993), *Associative Engines*. Cambridge, Mass: MIT Press.

Clark, A. (1996), 'Dealing in Futures: Folk Psychology and the Role of Representations in

Cognitive Science', in R. McCauley (ed.), *The Churchlands and Their Critics*. Cambridge, Mass: Blackwell.

Clark, A. (1997), *Being There: Putting Brain, Body and World Together Again*. Cambridge, Mass: MIT Press.

Clark, A. (2001), *Mindware*. Oxford: O.U.P.

Clark, A. and C. Thornton (1997), 'Trading Spaces: Computation, Representation and the Limits of Uninformed Learning', *Behavioural and Brain Sciences* 20(1), pp. 57-92.

Cohen, J. & F. Tong (2001), *Science* 293, pp. 2405-2407

Cowie, F. (1999). *What's Within?* (Oxford, OUP).

Cummins, R. (1984). 'Functional Analysis'. *Conceptual Issues in Evolutionary Biology: An Anthology*. E. Sober. (ed.) (Cambridge, Mass., Bradford, MIT Press).

Cummins, R. (1989). *Meaning and Mental Representation*. (Cambridge, MA, MIT Press).

Cummins, R. (1996), *Representations, Targets, and Attitudes*. Cambridge, Mass: MIT Press.

Davidson, D. (1974). 'Belief and the Basis of Meaning.' *Synthese* 27: 309-23.

Davidson, D. (1980), *Inquiries into Truth and Interpretation*. Oxford: O.U.P.

Davidson, D. (1987). 'Knowing One's Own Mind.' *The Proceedings and Address of the A.P.A.* 60: 441-58.

Davidson, D. (1993), 'Thinking Causes' in J. Heil and A. Mele (eds.), *Mental Causation*. Oxford: Clarendon Press.

Davidson, D. (1995), 'Could There Be A Science of Rationality?', *International Journal of Philosophical Studies* 3(1), pp. 1-16.

Davies, M. (1991), 'Individualism and Perceptual Content', *Mind* 100, pp. 461-484.

Dawson, M. and C. D. Piercey (2001), 'On the Subsymbolic Nature of a PDP Architecture that Uses a Nonmontonic Activation Function', *Minds and Machines* 11, pp. 197-218.

Dawson, M. et al (2000), 'Using extra output learning to insert a symbolic theory into a connectionist network', *Minds and Machines* 10, pp. 171-201.

Dennett, D. (1978), *Brainstorms*. Cambridge, Mass: MIT Press.

Dennett, D. (1981a). 'A Cure for the Common Code?' *Brainstorms: Philosophical Essays on Mind and Psychology*. (ed.) (Brighton, Harvester Press).

Dennett, D. (1981b). 'True Believers: The Intentional Strategy and Why It Works'. *Scientific Explanation*. A. F. Heath. (ed.) (Oxford, O.U.P.).

Dennett, D. (1987). *The Intentional Stance*. (Cambridge, MA, MIT Press).

Dennett, D. (1991). 'Real Patterns.' *Journal of Philosophy* 88: 27-51.

Dickinson, A. (1994). 'Instrumental Conditioning'. *Animal learning and cognition*. N. J. Mackintosh. (ed.) (San Diego, CA, Academic Press): 45-79.

Dietrich, E. and A. Markman (2003). 'Discrete Thoughts: Why Cognition Must Use Discrete Representations.' *Mind & Language* 18(1): 95-119.

Downing, P., et al (2001), *Science* 293, pp. 2470-2473

Dretske, F. (1977). 'Laws of Nature.' *Philosophy of Science* 44: 248-68.

Dretske, F. (1981), *Knowledge and the Flow of Information*. Cambridge, Mass: MIT Press.

Dretske, F. (1986), 'Misrepresentation' in *Belief: Form, Content and Function*, R. J. Bogdan (ed.), Oxford: O.U.P. Repr. in *Mental Representation: A Reader,* S. P. Stich & T. A. Warfield (eds.) 1994. Oxford: Blackwell.

Dretske, F. (1991). *Explaining Behaviour: Reasons in a World of Causes*. Cambridge, Mass: MIT Press.

Dummett, M. (1976). 'What is a Theory of Meaning? (II)'. *Truth and Meaning: Essays in Semantics*. G. Evans and J. McDowell. (ed.) (Oxford, OUP).

Egan, F. (1991), 'Must Psychology be Individualistic?', *Philosophical Review* (1991), pp. 179-203.

Egan, F. (1992), 'Individualism, Computation & Perceptual Content', *Mind* 101, pp. 443-459.

Elman, J. (1990), 'Finding structure in time', *Cognitive Science* 14, pp. 179-212.

Elman, J. (1991), Incremental Learning, or the Importance of Starting Small. Technical report 9101, Center for Research in Language, U.C.S.D. Described in detail in Clark (1993), pp. 138-142.

Elman, J. (1991a), 'Distributed representations, simple recurrent networks, and grammatical structure', *Machine Learning* 7, pp. 195-225.

Elman, J. (1992), 'Grammatical structure and distributed representations', in S. Davis (ed.), *Connectionism: theory and practice*. New York: O.U.P.

Elman, J. L., E. A. Bates, et al. (1996). *Rethinking Innateness*. (Cambridge, MA, MIT Press).

Field, H. (1978), 'Mental Representation', *Erkenntnis* 13, pp. 9-61. Repr. in *Mental Representation: A Reader,* S. P. Stich & T. A. Warfield (eds.) 1994. Oxford: Blackwell.

Fodor, J. (1975). *The Language of Thought*. (Cambridge, Harvard University Press).

Fodor, J. (1980), 'Methodological solipsism considered as a research strategy in cognitive psychology', *Behavioural and Brain Sciences* 3, pp. 63-73.

Fodor, J. (1981). 'The Present Status of the Innateness Controversy'. in his *Representations: Philosophical Essays on the Foundations of Cognitive Science*. (Cambridge, MA, MIT Press): 257-316.

Fodor, J. (1983). *The Modularity of Mind* (MIT Press).

Fodor, J. (1985). 'Précis of *The Modularity of Mind*.' *Behavioural and Brain Sciences* 8.

Fodor, J. (1987). *Psychosemantics* (MIT Press).

Fodor, J. (1990). *A Theory of Content and Other Essays*. (Cambridge, MA, MIT Press).

Fodor, J. (1994), *The Elm and the Expert*. Cambridge, Mass: MIT Press.

Fodor, J. (1998), *Concepts: Where Cognitive Science Went Wrong*. Oxford: Clarendon Press.

Fodor, J. & E. Lepore (1992), *Holism: A shopper's guide*. Oxford: Blackwell.

Fodor, J. & E. Lepore (1993), 'Reply to Churchland', *Philosophy and Phenomenological Research* 53, pp. 679-683.

Fodor, J. and E. Lepore (1996). 'The Red Herring and Pet Fish: Why Concepts Still Can't Be Prototypes.' *Cognition* 58: 253-270.

Fodor, J. & E. Lepore (1999), 'All at sea in semantic space: Churchland on meaning similarity', *Journal of Philosophy* 96(8), 381-403.

Fodor, J. A. and B. McLaughlin (1990). 'Connectionism and the problem of systematicity:

Why Smolensky's solution doesn't work.' *Cognition* 35: 183-204.

Fodor, J. A. and Z. W. Pylyshyn (1988). 'Connectionism and Cognitive Architecture: A Critical Analysis.' *Cognition* 28: 3-71.

Gelman, S. A., J. Coley and G. Gottfried (1994). 'Essentialist beliefs in children: the acquisition of concepts and theories'. *Mapping the Mind*. L. Hirschfeld and S. A. Gelman. (ed.) (Cambridge, CUP): 341-365.

Gelman, S. and J. D. Coley (1991). 'Language and categorization: the acquisition of natural kind terms'. *Perspectives on language and thought: interrelations in development*. S. Gelman and J. P. Byrnes. (ed.) (Cambridge, CUP).

Gelman, S. and L. Hirchfield (1999). 'How Biological is Essentialism?' *Folkbiology*. D. Medin and S. Atran. (ed.) (Cambridge, MA, MIT Press).

Godfrey-Smith, P. (1994). 'A Continuum of Semantic Optimism'. *Mental Representation: A Reader*. S. P. Stich and T. A. Warfield. (ed.) (Oxford, Blackwell).

Goldstone, R. & B. Rogosky (2002), 'Using relations within conceptual systems to translate across conceptual systems', *Cognition* 84, pp. 295-320.

Goodman, N. (1955). *Fact, Fiction and Forecast*. (London, Univ. of London Press).

Griffiths, P. and R. Gray (1992). 'Developmental Systems and Evolutionary Explanations.' *Journal of Philosophy* 91: 277-304.

Grindley, G. C. (1932). 'The formation of a simple habit in guinea pigs.' *British Journal of Psychology* 23: 127-147.

Hampton, J. (1987). 'Inheritance of Attributes in Natural-Concept Conjunctions.' *Memory and Cognition* 15: 55-71.

Haxby, J., et al (2001), *Science* 293, pp. 2425-2430

Haybron, D. (2000), 'The Causal and Explanatory Role of Information Stored in Connectionist Networks', *Minds and Machines* 10, pp. 361-380.

Hempel, C. G. (1965). *Aspects of Scientific Explanation and Other Essays*. (New York, Free Press).

Hinton, G. (1989), 'Connectionist learning procedures', *Artificial Intelligence* 40, pp. 185-234.

Hornsby, J. (1997). *Simple Mindedness: A Defence of Naïve Naturalism in the Philosophy of Mind*. (Cambridge, MA, Harvard University Press).

Huntley-Fenner, G., S. Carey and A. Solimando (2002). 'Objects are individuals but stuff doesn't count: perceived rigidity and cohesiveness influence infants' representation of small groups of discrete entities.' *Cognition* 85: 203-221.

Hurford, J. (2002). 'The Neural Basis of Predicate-Argument Structure.' *Behavioural and Brain Sciences*.

Ishai, A., L. G. Ungerleider, A. Martin, et al (1999). 'Distributed Representation of Objects in the Human Ventral Visual Pathway.' *Proceedings of the National Academy of Sciences* 96: 9379-9384.

Jackendoff, R. (1989). 'What Is a Concept, That a Person May Grasp It?' *Mind & Language* 4: 68-102.

Jackson, F. & P. Pettit (1993), 'Some Content is Narrow' in J. Heil and A. Mele (eds.), *Mental Causation*. Oxford: Clarendon Press.

Kandel, E. R. and R. D. Hawkins (1992). 'The Biological Basis of Learning and Individuality.' *Scientific American* 267: 62-71.

Kanwisher, N. (2000), 'Domain specificity in face perception', *Nature Neuroscience* 3, pp.

759 – 763.

Karmiloff-Smith, A. (1994), 'Précis of: *Beyond Modularity: A developmental perspective*', *Behavioural and Brain Sciences* 17(4), pp. 693-745.

Keil, F. C. (1989). *Concepts, Kinds and Cognitive Development*. (Cambridge, MA, MIT Press).

Keil, F. C. and N. Batterman (1984). 'A characteristic-to-defining shift in the development of word meaning.' *Journal of Verbal Learning and Verbal Behaviour* 23: 221-236.

Kim, J. (1998), *Mind in a Physical World*. Cambridge, Mass: MIT Press.

Kripke, S. (1972), *Naming and Necessity* (Oxford, Blackwell).

Laakso, A. & G. Cottrell (2000), 'Content and cluster analysis: assessing representational similarity in neural systems', *Philosophical Psychology* 13(1), 47-76.

Lakoff, G. (1972), 'Hedges: A study in meaning criteria and the logic of fuzzy concepts' in *Papers from the eighth regional meeting, Chicago Linguistics Society*, (Chicago, University of Chicago Linguistics Department).

Lakoff, G. (1987), 'Cognitive Models and Prototype Theory' in *Concepts and Conceptual Development: Ecological and Intellectual Factors in Categorisation,* U. Neisser (ed.), (Cambridge, CUP), pp. 63-100.

Landau, B. (1982). 'Will the real grandmother please stand up.' *J. Psycholing. Res.* 11(2): 47-62.

Laurence, S. and E. Margolis (1999), 'Concepts and Cognitive Science' in *Concepts: Core Readings,* E. Margolis and S. Laurence (eds.), (Cambridge, MA, MIT Press).

Laurence, S. and E. Margolis (2002a). 'Concepts'. *The Blackwell Guide to the Philosophy of Mind*. S. Stich and T. Warfield. (ed.) (Oxford, Blackwell): 190-213.

Laurence, S. and E. Margolis (2002b). 'Radical Concept Nativism.' *Cognition* 86: 25-55.

Lehky, S. and T. Sejnowski (1987), 'Extracting 3-D curvatures from images using a neural model', *Society for Neuroscience Abstracts* 13, 1451.

Lehky, S. and T. Sejnowski (1988), 'Neural network model for the representation of surface curvature from images of shaded surfaces', in J. Lund (ed.) *Organising Principles of Sensory Processing*. Oxford: O.U.P.

Lettvin, J., et al (1959), 'What the frog's eye tells the frog's brain', *Proceedings of the Institute of Radio Engineers* 47, pp. 1940-1957.

Lewis, D. (1970). 'How to Define Theoretical Terms.' *Journal of Philosophy* 67: 427-46.

Mandler, J. (1994). 'How to build a baby, II: conceptual primitives.' *Psychological Review* 99: 587-604.

Mandler, J. (2002). The Infant as Parent to the Adult. Conceptual Knowledge: Developmental, Biological, Functional and Computational Accounts, The British Academy, London.

Mandler, J. M. and L. McDonough (1998). 'Studies in inductive inference in infancy.' *Cognitive Psychology* 37: 60-96.

Maguire, E. A., N. Burgess, et al. (1998). 'Knowing Where and Getting There: A Human Navigational Network.' *Science* 280: 921-924.

Marr, D. (1982). *Vision*. (New York, W H Freeman & Co.).

Martin, A. (2002). Distributed Representation of Object Concepts in the Brain. Conceptual Knowledge: Developmental, Biological, Functional and Computational Accounts, The British Academy, London.

McClelland, J., D. Rumelhart, and the PDP Research Group (1986), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 2: Psychological and Biological Models*. Cambridge, Mass: MIT Press.

McLeod, P., K. Plunkett and E. Rolls (1998). *Introduction to Connectionist Modelling of Cognitive Processes*. (Oxford, OUP).

Medin, D. L. and M. M. Schaffer (1978). 'A context theory of classification learning.' *Psychological Review* 85: 207-238.

Millikan, R. (1984), *Language, Thought and Other Biological Categories*. Cambridge, Mass: MIT Press.

Millikan, R. (1989), 'Biosemantics', *Journal of Philosophy* 86, pp. 281-297. Repr. in *White Queen Psychology and Other Essays for Alice* 1993. Cambridge, Mass: MIT Press.

Millikan, R. (1996a). 'On Swampkinds.' *Mind & Language* 11(1): 103 - 117.

Millikan, R. (1996b), 'Pushmi-pullyu Representations', in James Tomberlin, ed., *Philosophical Perspectives* vol. IX, (Ridgeview Publishing) pp. 185-200. Reprinted in *Mind and Morals*, 1996, L. May and M. Friedman (eds.) (Cambridge MA, MIT Press), pp. 145-161.

Millikan, R. (2000), *On Clear and Confused Ideas* (Cambridge, Cambridge University Press).

Millikan, R. (2002), 'Biofunctions: Two Paradigms' in Cummins, Ariew and Perlman (eds.), *Functions: New Readings in the Philosophy of Psychology and Biology*. Oxford: O.U.P.

Millikan, R. (manuscript). Some Kantian Reflections on Animal Minds. Oral presentation at conference: The Evolution of the Intelligent Mind, King's College, London, 15th April 2000. Found as an unpublished manuscript under the title 'Some Different Ways to Think'

Milner, A. D. and M. A. Goodale (1995). *The Visual Brain in Action*. (Oxford, OUP).

Mozer, M. and P. Smolensky (1989), 'Using relevance to reduce network size automatically', *Connection Science* 1(1), pp. 3-17.

Murphy, G. and D. Medin (1985). 'The Role of Theories in Conceptual Coherence.' *Psychological Review* 92(3): 289-316.

Neander, K. (1996). 'Swampman Meets Swampcow.' *Mind & Language* 11(1): 118 - 129.

O'Keefe, J. and N. Burgess (1996), 'Geometric determinants of the place fields of hippocampal neurons', *Nature* 381, pp. 425-428.

O'Keefe, J. and L. Nadel (1978). *The hippocampus as a cognitive map*. (Oxford, Clarendon Press).

Osherson, D. and E. Smith (1981). 'On the Adequacy of Prototype Theory as a Theory of Concepts.' *Cognition* 9: 35-58.

Papineau, D. (1987), *Reality and Representation*. Oxford: Blackwell.

Papineau, D. (1993), *Philosophical Naturalism*. Oxford: Blackwell.

Papineau, D. (1996). 'Doubtful Intuitions.' *Mind & Language* 11: 130-132.

Papineau, D. (2001). 'The Status of Teleosemantics, or How to Stop Worrying About Swampman.' *Australasian Journal of Philosophy* 79(2): 279-289.

Peacocke, C. (1983). *Sense and Content: Experience, Thought and Their Relations*. (Oxford, OUP).

Peacocke, C. (1992), *A Study of Concepts* (Cambridge, MA, MIT Press).

Pearce, J. M. (1997). *Animal Learning and Cognition*. (Hove, Psychology Press).

Pollack, J. (1990), 'Recursive distributed representations', *Artificial Intelligence* 46, pp. 77-105.

Price, C. (2001). *Functions in Mind*. (Oxford, Clarendon Press).

Prinz, J. (2002). *Furnishing the Mind*. (Cambridge, MA, MIT Press).

Putnam, H. (1970), 'Is Semantics Possible?' in *Language, Belief and Metaphysics,* H. Kiefer and M. Munitz (eds.), (New York, State University of New York Press)**:** 50-63. Repr. in *Concepts: Core Readings*, E. Margolis and S. Laurence (eds.) 1999 (Cambridge, MA, MIT Press).

Putnam, H. (1975). 'The Meaning of "Meaning"', repr. in his *Mind, Language and Reality*, 1979 (Cambridge, CUP).

Quine, W. V. (1970). 'Natural Kinds'. *Essays in Honor of Carl G. Hempel*. N. Rescher. (ed.) (Dordrecht, D. Reidel)**:** 1-23.

Ramsey, W., S. Stich and J. Garon (1990). 'Connectionism, Eliminativism, and the Future of Folk Psychology.' *Philosophical Perspectives* 4: 499-533. Repr. in *Connectionism: Debates in Psychological Explanation*, C. MacDonald and G. MacDonald (eds.), 1995 (Oxford: Blackwell).

Reed, S. K. (1972). 'Pattern recognition and categorisation.' *Cognitive Psychology* 3: 382-407.

Rey, G. (1983). 'Concepts and Stereotypes.' *Cognition* 15: 237-262.

Rey, G. (1994), 'Concepts' in *A Companion to the Philosophy of Mind,* S. Guttenplan (ed.), (Oxford, Blackwell).

Ridley, M. (2003). *Nature Through Nurture*.

Rips, L. J. (1989). 'Similarity, typicality, and categorization'. *Similarity in analogical reasoning*. S. Vosniadou and A. Ortony. (ed.) (Cambridge, CUP): 21-59.

Rolls, E. and A. Treves (1998). *Neural Networks and Brain Function*. (Oxford, OUP).

Rosch, E. (1974), 'Linguistic Relativity' in *Human Communication: Theoretical Perspectives*, A. Silverstein (ed.), (New York, Halsted Press).

Rosch, E. (1975). 'Cognitive Representations of Semantic Categories.' *Journal of Experimental Psychology: General* 104: 192-233.

Rosch, E. (1977), 'Human categorization' in *Advances in cross-cultural psychology*, N. Warren (ed.), (London, Academic Press).

Rosch, E. (1978). 'Principles of Categorisation'. *Cognition and Categorisation*. E. Rosch and B. Lloyd. (ed.) (Hillsdale, NJ, Laurence Erlbaum).

Rosch, E. and C. B. Mervis (1975). 'Family Resemblances: Studies in the Internal Structure of Categories.' *Cognitive Psychology* 7: 573-605.

Rosch, E., C. Mervis, et al (1976). 'Basic objects in natural categories.' *Cognitive Psychology* 8(382-439).

Rosch, E., C. Simpson, et al (1976). 'Structural bases of typicality effects.' *Journal of Experimental Psychology: Human perception and Performance* 2: 491-502.

Rose, S. (1992). *The Making of Memory*. (London, Bantam Press).

Rumelhart, D. (1989), 'The Architecture of Mind: A Connectionist Approach', in M. Posner (ed.), *Foundations of Cognitive Science*. Cambridge, Mass: MIT Press. Repr. in J. Haugeland (ed.), *Mind Design II*, 1997. Cambridge, Mass: MIT Press.

Rumelhart, D., J. McClelland, and the PDP Research Group (1986), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 1:*

Foundations. Cambridge, Mass: MIT Press.

Rumelhart, D., P. Smolensky, J. McClelland and G. Hinton (1986), 'Schemata and sequential thought processes in PDP models' in McClelland & Rumelhart (1986).

Rupert, R. (2001), 'Coining Terms in the Language of Thought', *Journal of Philosophy,* Oct 2001, pp. 499-530.

Ryle, G. (1949), *The Concept of Mind*. London: Hutchinson & Company.

Sacks, O. (1985). *The Man Who Mistook His Wife for a Hat*. (New York, Summit Books).

Samuels, R. (2002), 'Nativism in Cognitive Science', *Mind and Language* 17(1), pp. 233-265

Segal, G. (1989), 'Seeing What is Not There', *Philosophical Review* 98, pp. 189-214.

Segal, G. (1991), 'In Defence of a Reasonable Individualism', *Mind* 100, pp. 485-494.

Segal, G. (2000). *A Slim Book About Narrow Content*. (Cambridge, MA, MIT Press).

Segal, G. (2003). 'Intentionality'. *Oxford Companion to Contemporary Analytic Philosophy*. F. Jackson and P. Pettit. (ed.) (Oxford, OUP).

Segal, G. (forthcoming), 'Reference, Causal Powers, Externalist Intuitions and Unicorns'.

Sejnowski, T. and C. Rosenberg (1987), 'Parallel Networks that Learn to Pronounce English Text', *Complex Systems* 1, pp. 145-168.

Shepherd, G. M. (1994). *Neurobiology, 3rd edition*. (Oxford, OUP).

Smith, E. (1995), 'Concepts and Categorisation' in *Thinking: An Invitation to Cognitive Science, Volume 3,* E. Smith and D. Osherson (eds.), (Cambridge, MA, Harvard University Press): 3-33.

Smith, E. and D. Medin (1981), 'The Exemplar View' in *Categories and Concepts*, (Cambridge, MA, Harvard University Press).

Smith, E., D. Osherson, et al (1988). 'Combining Prototypes; A Selective Modification Model.' *Cognitive Science* 12: 485-527.

Smolensky, P. (1986), 'Neural and conceptual interpretation of PDP models', in Rumelhart & McClennan (1986).

Smolensky, P. (1988). 'On the Proper Treatment of Connectionism.' *Behavioural and Brain Sciences* 11: 1-74.

Smolensky, P. (1991). 'Connectionism, Constituency and the Language of Thought'. *Meaning in Mind*. B. Loewer and G. Rey. (ed.) (Oxford, Blackwell).

Smolensky, P. (1995). 'Constituent Structure and Explanation in an Integrated Connectionist/Symbolic Cognitive Architecture'. *Connectionism: Debates on Psychological Explanation, Vol. 2*. C. MacDonald and G. MacDonald. (ed.) (Oxford, Blackwell).

Soja, N., S. Carey and E. Spelke (1991). 'Ontological categories guide young children's inductions on word meaning: object terms and substance terms.' *Cognition* 38: 179-211.

Steels, L. (1997). 'Synthesising the origins of language and meaning using co-evolution, self-organisation and level formation'. *Evolution of Human Language*. J. Hurford, C. Knight and M. Studdert-Kennedy. (ed.) (Edinburgh, Edinburgh University Press).

Steels, L. and B. De Boer (1996), 'Experiments in Emergent Phonology', VUB, Belgium; available online; arti.vub.ac.be/steels/publications.htm

Sterelny, K. (1990). *The Representational Theory of Mind: An Introduction*. (Oxford, Blackwell).

Stich, S. P. (1983). *Folk Psychology and Cognitive Science: The Case Against Belief*. (Cambridge, MA, MIT Press).

Strevens, M. (2000). 'The essentialist aspect of naïve theories.' *Cognition* 74: 149-175.

Sturgeon, S. (1998), 'Humean Chance: Five Questions for David Lewis', *Erkenntnis* 49, pp. 321-335.

Tiffany, E. (1999), 'Semantics San Diego Style', *Journal of Philosophy* 96, pp. 416-429.

Tversky, A. (1977). 'Features of Similarity.' *Psychological Review* 84: 327-352.

Tye, M. (1998), 'Inverted Earth, Swampman, and Representationalism', pp. 459-478 of *Philosophical Perspectives, 12, Language, Mind and Ontology,* J. E. Tomberlin (ed.), (Blackwell: Oxford).

Ungerleider, L. G. and M. Mishkin (1982). 'Two Cortical Visual Systems'. *Analysis of Visual Behavior*. D. J. Ingle, M. A. Goodale and R. W. J. Mansfield. (ed.) (Cambridge, MA, MIT Press).

Usher, M. (2001). 'A Statistical Referential Theory of Content: Using Information Theory to Account for Misrepresentation.' *Mind & Language* 16(3): 311-334.

Webb, B. (1994), 'Robotic Experiments in Cricket Phonotaxis', in *From Animals to Animats 3: Proceedings of the Third International Conference on Simulation of Adaptive Behaviour*, D. Cliff, P. Husbands, J. Meyer and S. Wilson (eds.), pp. 45-54.

Werning, M. (2003). 'The Temporal Dimension of Thought.' *Synthese*: forthcoming 2003.

Wheeler, M. and A. Clark (1999), 'Genic Representation: Reconciling Content and Causal Complexity', *British Journal for the Philosophy of Science* 50, pp. 103-135.

Whyte, J. (1990), 'Success Semantics', *Analysis* 50, pp. 149-157.