Neural Computation and Symbolic Thought

The Philosophical Implications of Biological Plausibility in Connectionist Network Modelling

Robert G. West

Contents

Introduction 1

1 Connectionism and Neural Computation 6

1.1 Connectionist Networks 6

1.2 Vector Cognition 17

1.2.1 Vector Coding 17

1.2.2 Vector Prototypes and Explanatory Understanding 19

1.3 Connectionism and Systematicity 25

1.4 Summary 29

2 Biologically Plausible Neural Computation 31

2.1 Pattern Association Networks 33

2.2 Autoassociation Networks 34

2.3 Competitive Networks and Convergent Architectures 35

2.4 Interactions between Neuronal Populations and Brain Function 41

2.4.1 Orthogonalization, Emotion, and Memory 42

2.4.2 Sparsification 44

2.4.3 Recurrent Connections and Neural Processing 45

2.5 General Principles of Brain Function 46

2.6 Summary 50

3 Rethinking Vector Cognition 51

3.1 Icon, Index, and Symbol 51

3.2 Neural Networks and Symbols 55

3.3 Symbols and Semantic Content 59

3.3.1 Context and Meaning 59

3.3.2 Meaning and Indexical Foundations 62
3.4 Symbols, Systematicity, and Concepts 63
3.4.1 Neural Commitments 65
3.5 Summary 72
4 Conclusions and Further Work 73
Bibliography 75

Introduction

Few would question the claim that *some* psychological and neuropsychological findings can be relevant to philosophy, and the philosophy of mind in particular. Many philosophers, however, have underestimated the intimacy of the relationship for certain types of empirical research, including neurophysiology, connectionist modelling, artificial life, situated robotics, cognitive psychology, and developmental psychology. Principal amongst these is the investigation of real and artificial neural networks, which I will concentrate upon. The investigation of such networks is still in its infancy, and far from revealing how the brain works. But the view of cognition which it hints at is so suggestive, and so promising, given the problems faced in the philosophy of mind, that I think it is worth sketching out and exploring from a philosophical perspective. I will argue that an investigation of the basic modes of neural processing suggests a radical alteration of the conceptual framework within which we attempt to understand the mind and cognition. I will advocate an approach on which these empirically based concepts actually form an important source for *philosophical* theory, shaping the nature of the concepts employed at the philosophical level of inquiry.

This approach may sound highly reductive, but I hope to make it clear that this is not the case if connectionism is properly construed. Connectionism demonstrates that simple models can give rise to complex effects which can only be understood at a higher level of abstraction. The basic mechanisms of connectionist models can be described in a couple of brief paragraphs, but their behaviour is not so transparent, requiring complex and difficult analysis. I am struck by a parallel here with the science of self-organized criticality, which attempts to explain how complexity emerges in a universe governed by simple laws. Self-organized criticality has been discovered in a variety of situations from earthquakes through biological evolution to traffic jams; its key features can be illustrated using the example of a sandpile formed by a slow steady trickle of sand onto a platform of limited area. As the sandpile grows avalanches of sand occur. Eventually a steady state is reached in which the amount of sand leaving the pile is equal to the amount being added. At this

point a critical state is reached in which avalanches of all sizes occur, some local, some encompassing the whole pile. However, the sizes of the avalanches are not random, they follow a power law, so that there are very few large avalanches, and very many small ones, although there is no correlation in the sizes of the avalanches from one moment to the next, just as rolling a six on a die does not alter the probability of rolling another one with the next throw. A plot using a log scale on both axes of size against frequency reveals a straight line. Bak explains this as follows:

> The addition of grains of sand has transformed the system from a state in which the individual grains follow their own local dynamics to a critical state where the emergent dynamics are global. In the stationary [self-organized critical] state, there is one complex system, the sandpile, with its own emergent dynamics. The emergence of the sandpile could not have been anticipated fom the properties of the individual grains.[1]

In the self-organized critical state the effect of small perturbations anywhere in the system cannot be predicted unless one knows the state of the whole sandpile and has a supercomputer available. Bak also argues that the brain operates as a self-organized critical system. Whether this is the case or not remains to be seen, but what is clear from this example is that a system of simple interconnected elements following simple rules, either grains of sand or neurons, can produce a highly complex pattern of emergent behaviour, and this pattern is simply missed if one attempts a reductive analysis.

The notion of emergent properties is notoriously vague, and gives rise to accusations of mysticism, but I will argue that it can be placed on a sound footing. For surely, if even physicists are starting to see the need for a non-reductive approach, which involves analysis of complete systems, then it cannot be unreasonable for cognitive scientists to do the same. I do not mean to imply that emergent properties are not supervenient upon the activities of fundamental particles. To think this would be to misconstrue the point as an *ontological* one, but I am attempting to address the issue of *explanation*. For if one acknowledges the existence of emergent phenomena, one must find some way to analyse them, and the science of self-organized criticality suggests how this might be done. Self-organized criticality has been investigated

---

[1] *How Nature Works: The Science of Self-Organized Criticality* (Oxford University Press, Oxford, 1997), p.51.

using highly simplified models, which nevertheless exhibit the phenomenon in question. In this way the modelling process is valuable because the designer has complete control of all the model's variables, thus the model can stand as proof that even where only simple local rules are operating global emergent properties can be produced. This simplification also reveals the fundamental principles at work, allowing an understanding of them, even if this requires a systems level analysis. This simplifying ethos is also found amongst connectionist modellers, and rightly so. Connectionism's avowed intent is to uncover the fundamentals of brain processing rather than recreating in intricate detail the neurophysiological processes that actually occur in the brain. This has lead to many exciting and interesting discoveries, about such features as memory, for example. However, I will argue that in some aspects the drive towards simplification has gone too far, the baby has been thrown out with the bath water. The hard part of the task is knowing when to stop jettisoning details.

With this in mind, rather than discussing the finer points of experimental methodology, I will try to draw out the important lessons from the way that ensembles of connectionist units, or neurons, function in order to suggest how a model of cognition might be produced which is pitched at a more abstract and general level.[2] Thus the nature of the underlying processing shapes the nature of mental processes, without providing a complete account of them. This, I hope, will allow a reconciliation of empirical and philosophical approaches.

My own approach was prompted by the work of the Churchlands, although it diverges significantly in its philosophical conclusions. I will take their view of neural computation as a starting point, indicating its limitations, and suggesting how more careful attention to the workings of real brains might overcome them. Thus I make a distinction between typical connectionist models and more biologically plausible models. I will argue that attention to biological detail can have consequences for our understanding of the philosophy of mind, particularly of symbolic thought. This will involve a certain amount of armchair speculation about the nature of brain processing

[2] Many connectionist researchers use the term 'neuron' to refer to the computational elements in their networks, but I shall use the term 'unit' in this context, and reserve 'neuron' for actual biological neurons. This is purely a matter of terminology on my part in order to make a distinction between more and less biologically plausible networks — although there is no strict dividing point with biologically plausible models on one side and the rest on the other, rather there is a spectrum of increasing biological plausibility.

and so it is open to empirical refutation, and indeed I doubt what I have to say will be correct in every detail. However, I believe that this does not vitiate its worth, for it demonstrates how empirical findings can be interestingly relevant to the philosophy of mind. Thus it has value as an intellectual exercise, in the same way that speculation about historical events based upon counterfactual assumptions extends one's appreciation of the workings of the actual course of historical events.

One aspect of my argument will involve comparison with sentential models of cognition. For the motivation behind the neuronally inspired approach is the belief that, at base, cognition does not consist in sentence-crunching (i.e., operations involving symbols, as take place in a digital computer). There is no need to postulate a language of thought in order to explain the conceptual, combinatorial, and productive aspects of cognition. In attempting to explain these phenomena, advocates of connectionism attempt to demonstrate that such networks can model linguistic and rule-governed behaviour, such as past tense formation. Symbolic theorists cite the fragmentary and detached character of these examples as evidence that the connectionist style of processing will prove inadequate to the task of explaining the essentially systematic nature of human thought.[3] Thus until a fully-fledged language-using connectionist network is constructed the argument cannot be settled. I do not mean to be derogatory about the orthodox connectionist approach, indeed it has produced stimulating and valuable results. Yet if the argument is to be resolved then efforts must be made to understand how real neural systems utilize neurons, in terms of the configuration of their connections, to produce complex, including linguistic, behaviour.

Thus the challenge for connectionism, and cognitive science in general, is to try and understand how the great structural differentiation of the brain relates to its essentially distributed mode of processing and the properties of individual neurons. For it is not enough simply to keep trying to model high level tasks using simple individual networks and complex learning algorithms. Rather attempts should be made to model the total behaviour of basic organisms, building upwards in the hope

---

[3] See J. Fodor and Z. Pylyshyn, 'Connectionism and Cognitive Architecture: A Critical Analysis', in S. Pinker and M. Jacques, eds., *Connections and Symbols* (MIT Press, Cambridge, Mass., 1988), pp. 3-71, for the locus classicus of this type of attack against connectionism.

of understanding the role of the rich structure and modularity of neural systems in cognitive processing. In this way an understanding of how large groups of neurons interact to process highly complex information may allow artificial networks to exhibit ever more complex forms of behaviour. The beginnings of this task are already underway, and perhaps in this manner it might be possible to produce a fully-fledged concept-using parallel processing system at some point in the future.

However, I believe that we do not have to wait for this event before we can at least begin to ground the claim that language and concept-use can be understood in terms of a non-sentential form of processing.[4] I will argue that connectionism, as it stands, cannot adequately account for symbolic and conceptual thought, and that the elements of real neural processing from which connectionist models have abstracted away provide the raw materials with which it is possible to answer the critics who believe that parallel processing can never explain symbolic thought. In my view, some of the properties of higher cognitive functions will be best understood by reflecting, in detail, on the methods that real brains actually use. This analysis depends upon the large-scale principles that have so far been gleaned from the study of neural computation, combined with recent speculation about the nature of symbols. Thus, I am being somewhat peremptory, for at present large-scale principles of neuronal processing remain inchoate, but I think that enough is in place to warrant a first speculative survey.

[4] Indeed, such ruminations may help to shape the research strategy employed in uncovering the principles of large scale neuronal processing.

1 Connectionism and Neural Computation

In this chapter I first want to present some of the principles of large scale neural processing which have so far been gleaned from the study of connectionist networks. My intent is not to present evidence or argument for their validity — that is not a proper task for philosophy, and so should be left to others. Rather, my concern is to present their details so that their philosophical consequences, if they prove correct, can be explored. I then want to give a brief sketch of Churchland's neurocomputational perspective before arguing that it is in need of significant alteration.

1.1 Connectionist Networks

Connectionist networks can take many forms, but the features which they typically share are simple processing units, and weighted connections between those units. The activity of each unit is determined by the sum of the activation it receives from other units via its input connections. The strength of the input it receives from each of these units is determined by the product of the input unit's activity and the weight of the connection. All of these inputs are then summed, to give the net input, which is fed into the activation function, to determine the level of activity of that unit, which is then transmitted via outgoing connections to other units. The activation function can take a variety of forms, as displayed in figure 1 on page 7. An example is shown in Table 1, on page 8. The major distinction is between linear and non-linear functions, and the latter are most common because they allow more complex problems to be tackled. A prototypical connectionist network consists of three layers of units, connected in a feedforward fashion, so that each unit in a layer has a connection to every unit in the next layer. There is an input layer, an output layer, and a hidden layer, so-called because it is not directly connected to the external environment. Most of what follows concerns this type of architecture, as it demonstrates the majority of important features of connectionist networks.

Connectionist networks can utilize several types of representations, and these have considerable consequences for the behaviour and interpretation of the network.

The most important type is ***distributed representation***, where all of the units in a layer are involved in representing a given item in the task domain; the representational vehicle is the pattern of activity. In all of the networks discussed below the hidden layers, at the very least, have distributed representations. Representations can be more or less distributed. For simple binary units, which are either on or off (0 or 1) if half of the units are on and the other half are off, e.g. 010011, then the representation is ***fully*** distributed. The reason for this epithet is that in order to know what is being represented one must know the activity of all six units. At the other end of the spectrum, representations are ***local*** if a stimulus is indicated by the activity of only one unit, as in 010000. Between these two extremes are vectors where a proportion of the units are involved in the representation of a stimulus, e.g. 010100, and this is called ***sparse*** representation.

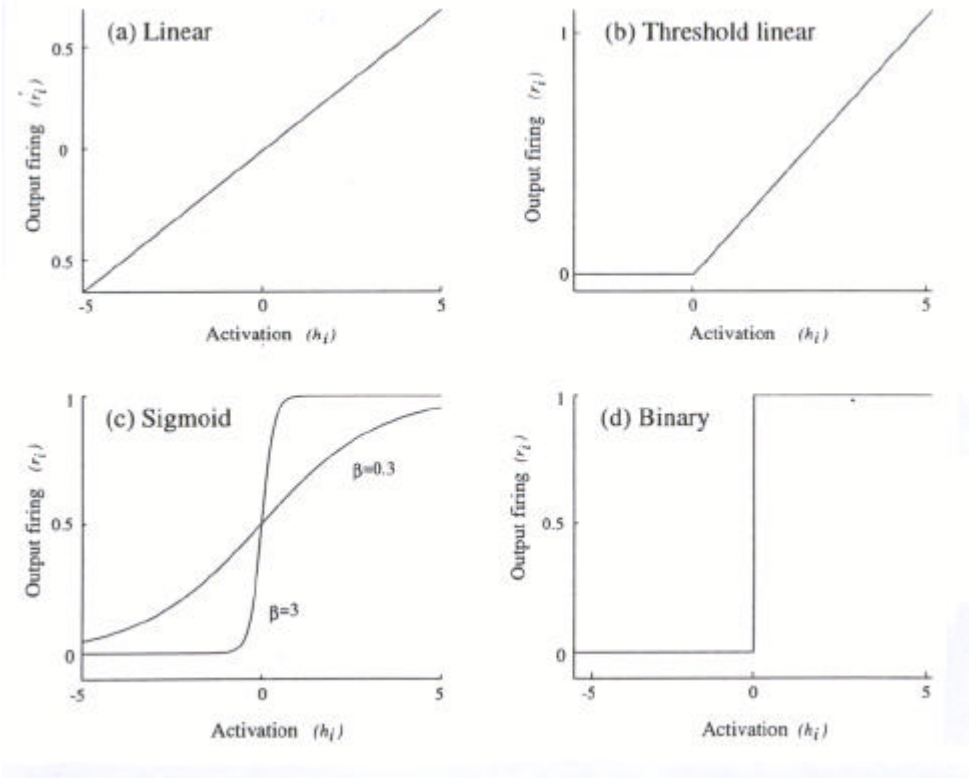

Figure 1 Different types of activation function. The activation function determines the output of a unit given its net input. (a) Linear. (b) Threshold Linear. (c) Sigmoid (the steepnes of the slope depends upon the exact nature of the function). (d) Binary threshold.

The representational primitives in feedforward networks are patterns of activation in a layer of units. These can be considered as vectors, where a vector is just defined as an ordered set of numbers. Vectors can be conceptualized

geometrically in terms of multidimensional spaces. Each unit in the population determines an axis of the space, and its level of activation specifies a point on that axis. Thus a given pattern of activation can be considered as a point in that space. This is most clearly appreciated for the input and output layers, where a pattern of activity could represent a sound wave, a digitized picture of a face, or a phonemic representation of a word. Similarly the output pattern of a network could be a real world action, such as the input to a speech synthesizer. Interpretation of the hidden layer of connectionist networks is not so easy. Such a layer is needed in order for networks to learn complex tasks successfully, and so an understanding of the way in which networks learn is needed in order to understand the role of the hidden layer.

One way to begin to understand the processing in networks is to view them as carrying out transformations from one vector to another. Vector transformation does not require input and output vectors to have the same dimensions, and so vastly differing representations, such as a sensory vector and a motor vector can be co-ordinated in a principled fashion. The example in Table 1 shows the transformation of a three-dimensional vector into a four-dimensional vector by a matrix.

		Weight Matrix			
Input Vector	0.6	0.6	3.2	6.0	-2.4
	0.3	-0.5	5.0	2.9	0.5
	0.4	1.5	-2.3	-1.2	-3.7
Net Inputs		0.81	2.5	3.1	-2.8
		0.69	0.92	0.96	0.06
		Output Vector			

Table 1: An example of vector transformation via a matrix.

To calculate the first term of the output vector, each term of the input vector is multiplied by the corresponding term in the matrix, and these are summed to give the net input: $(0.6 \cdot 0.6) + (0.3 \cdot -0.5) + (0.4 \cdot 1.5) = 0.81$. The net input is passed through an activation function to give the unit's activation, which can then be passed on to units in the next layer. The example in Table 1 involves units with activations that

vary between 0 and 1 according to the following sigmoid activation function, which is plotted in Figure 1(c), on page 7:

$$a_i = \frac{1}{1+e^{-net_i}}$$

where a_i is the unit's activation, and net_i is the net input of that unit. For any two vectors, there will always exist a tensor matrix which will produce the desired transformation between them. This in itself is unremarkable. The hard task is to explain how networks manage to orchestrate the connections between units so that the right transformations occur (where the right transitions are those which are behaviourally advantageous) for a whole range of inputs.

Initially random weights are assigned to a network's connections, and it is then presented with a set of training inputs. For each one the output of the network is compared with the correct output, the teacher pattern, so-called because it is externally determined, and the network's connections are minutely altered according to a learning algorithm. This calculates the error for a given unit, and then alters it connections by an amount relative to their influence in the production of the error. In this way the network's output is brought closer to the desired output. Through this process of error minimisation the network gradually comes to manifest the appropriate input/output mapping. In this process the only way in which the behaviour of the network is shaped is through the selection of the training set, the initial structure of the network, the activation function of the units, and the learning algorithm; although it is important to note that these are crucial in determining whether the network learns successfully or not.[5] During training the network is supervised in *some* sense, in that it requires an external error signal and learning algorithm to affect changes in its weights, but in a much more interesting sense the network functions autonomously, in that there is no programming. There is no reason why anyone, even the designer of the network, must understand how it managed the task. A simulation is set in motion and left to do its thing until (hopefully) it manages the assigned task. From this point of view connectionist networks appear to be as inscrutable as the brain. However, given that we have detailed information about the

[5] A good deal of pre-processing can be smuggled into the initial structure through the choice of input and output representations. Thus it is important that the nature of these representations be scrutinized in assessing the significance of any model.

patterns of activity and weight changes in such networks we can go about trying to explain and interpret the way in which networks manage to do what they do. One of the advantages of using computer models is the possibility of examining the innards of these systems in vivo, as it were, during their learning and subsequent operation.

There are two interrelated aspects of networks which are crucial to understanding their computational capacities, and both arise as a result of distributed representations: namely, their incorporation of a ***semantic metric***, and their use of ***superpositional storage***.[6] If representations involve a semantic metric this means that similarities in semantic content are reflected in similarities amongst representational vehicles. For example, one might have a layer of units which represent faces, forming a multidimensional vector space. Similar faces would be represented by points that are close together in this vector space. Further, the relationships between faces would also be reflected in the relative positions of their points in vector space. The midpoint on a line between two faces would appear similar to both, and a smooth progression along that line would appear as a gradual transformation from one of the faces to the other. A semantic metric is vitally important because it means that a network receives inputs which reflect the relations between the items represented. If genuine categories exist in the training set, then they will be present in the representations of that set. Thus similar faces will produce similar input vectors to a face recognition network. To see the implications of this for network processing one must understand the way in which networks store information, using superpositional storage.

A network's job, from the perspective of its designer, is to transform input vectors into the correct output vectors. Such mappings are achieved because the network's connection weights form a suitable matrix. So the connection weights are the repository of the network's experience of its training set. In the connectionist networks that I have been discussing this storage is superpositional, which means that exactly the same units are used to represent each item, because of the distributed nature of vector representation. Thus existing connection weights, encoding previous

---

[6] This notion was originally developed in T. van Gelder 'What is the 'D' in PDP'? A Survey of the Concept of Distribution', in W. M. Ramsey, D. E. Rumelhart, and S. P. Stich, eds., *Philosophy and Connectionist Theory* (Lawrence Erlbaum Associates, Hillsdale, N.J., 1991), pp. 33-59.

training experience, must be altered in order to accommodate further training input. Clark explains the consequences of this as follows:

> . . . semantic features which are statistically frequent in a body of input exemplars come to be both highly marked and mutually associated. By 'highly marked' I mean that the connection weights constituting the net's long-term stored knowledge about such common features tend to be quite strong, since the training regime has repeatedly pushed them to accommodate this pattern . . . By 'mutually associated' I mean that where such features co-occur, they will tend to become encoded in such a way that activation of the resources encoding one such feature will promote activation of the other. The joint effect of these two tendencies is a process of automatic prototype extraction: the network extracts the statistical central tendency of the various feature complexes and thus comes to encode information not just about specific exemplars but also about the stereotypical feature-set displayed in the training data.[7]

Thus networks latch onto statistical tendencies in their training set in order to produce the correct mapping. This can be stated in terms of vectors by saying that if there is a group of vectors in the training set which cluster in a region of vector space, then the network will come to treat them in the same way. This can be understood by looking at the properties of the dot product of an input activation vector and the weight vector for a given unit. These are just ordered lists of the values of the weights and activations that the unit receives. Assuming that the unit only receives one connection from each unit in the previous layer, the activation vector and the weight vector will have the same dimensions. Thus the weight vector will indicate the input activation pattern which would produce the largest activation in the receiving unit. Where a weight is large and positive, if the activation of its input unit is maximally positive (assuming bipolar units with activations ranging between –1 and +1) then the activation which is transmitted will be maximized. Hence one way to interpret the unit's behaviour is in terms of vector comparison. The weight vector for a unit indicates its preferred stimulus, and each input is compared with it. If the activation function of the unit is non-linear, i.e. if it has a firing threshold of some kind, then this will determine a criterion of similarity. Some patterns will be close enough to the ideal to push the unit past its threshold, and into firing, others will not, and the unit will remain inactive, or at a low level of activity if its activation function is sigmoid.

---

[7] *Associative Engines* (MIT Press, Cambridge, Mass., 1993), pp. 20-1.

However, this is only the first move in understanding how connectionist networks function. For it is not always a benefit for similar inputs to be treated in the same way. It is one of the advantages of connectionist networks that they can learn to place highly abstract boundaries across multidimensional similarity spaces, i.e. they can latch onto the right similarities, ignoring those that conflict with the categorisation task at hand. This ability requires a hidden layer, a point that can be demonstrated by considering the task of mapping the exclusive OR function, where the required transformations are as follows:

Input 1	Input 2	Required Output
0	0	0
1	0	1
0	1	1
1	1	0

Table 2: Mappings required for the exclusive OR problem.

A network without a hidden layer cannot solve this problem because the patterns are not linearly separable; i.e. no hyperplane can be placed in their N-dimensional space (where N is the number of units) so as to separate the input patterns requiring different responses. This is illustrated in Figure 2 (a) where it can easily be seen that no straight line can be drawn which has [0,1] and [1,0] on one side and [0,0] and [1,1] on the other. Points that are closer together, and therefore most similar to each other, must be treated differently to those which are further apart.

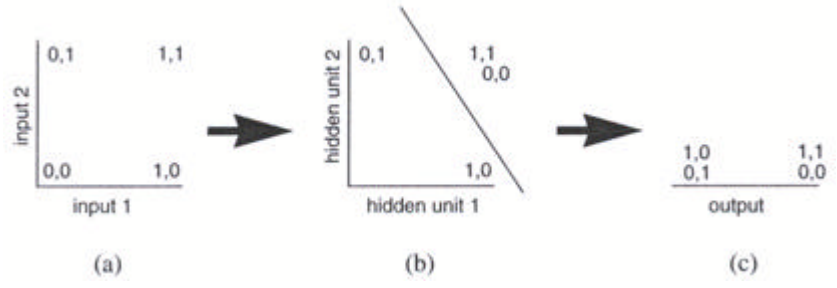

Figure 2: Vector space representation of the exclusive OR problem. In (a) the input space is shown, and it can be seen that no hyperplane can separate out the right points. In (b) the hidden unit vector space is shown with the points transformed in vector space so that a hyperplane can perform the separation, which is shown in (c).

This idea of a set of weights imposing a hyperplane is extremely important in thinking about the behaviour of networks, as it allows one to visualize what the network is doing. The hidden layer allows this problem to be overcome because it performs a transformation on the vector space, so that in the hidden unit vector space the originally dissimilar inputs are moved closer together, so that a hyperplane can separate them, as shown in Figure 2 (b) and (c). This separation is carried out by the connections to the output layer, where the solution is achieved.

At a more abstract level these processes of vector comparison and vector space transformation can be seen as the basis for prototype extraction, as Clark suggested in the passage quoted above. A word of caution is needed here, for if there are too many units in the hidden layer the network will merely learn a separate hidden layer activation pattern for each input in the training set. In other words it will not take advantage of the distributed representation in the input layer, as there is no need for it to use the same resources to store all the inputs patterns, so there is no superpositional storage, and no semantic metric. As a result when an input which was not amongst the training set is presented the network cannot produce an appropriate response. However, when there are not enough hidden units for this to occur, the network has to use its limited computational resources to successfully grasp the relevant similarities between training inputs, which will allow it to accurately categorize novel stimuli.[8]

As a result of training the network comes to funnel inputs into specific regions of hidden layer activation space. These regions in turn are recognized by the output layer as indicating a specific response. To give a specific — and hackneyed — example, Gorman and Sejnowski designed a network to distinguish the difference between sonar echoes from rocks and mines.[9] The frequency profile of the echo was represented on the input layer, and the output layer contained two units to indicate the network's decision, either rock or mine. Upon analysing its hidden layer they found it to be partitioned into two regions, activity in one caused the rock output unit to fire, and activity in the other causing the mine output unit to fire. Thus the hidden layer

---

[8] At present the only way to calculate the optimum number of hidden units is by rule of thumb, or trial and error

[9] 'Learned Classification of Sonar Targets Using a Massively-Parallel Network', *IEEE Transactions: Acoustics, Speech, and Signal Processing* (1988), 1135-1140.

activation space contained a similarity gradient moving from peaks at the central points of each region, which were most easily recognized by the output units, to the border between them, which produced an indecisive response from the output units. Thus these 'hotspots' can be considered as prototypes, as any input vector which is transformed into that region of hidden layer vector space would produce a response suitable for a given category, such as a mine. Hence network training can be seen as the alteration of connection weights so that input vectors are pushed into specific regions of the hidden layer vector space. The output layer then has the job of correctly recognising which of these region goes with which response. Whilst networks exhibit an ability to map previously experienced stimuli onto the correct prototype in this way, they also demonstrate an ability to generalize from experience to map a novel input onto the appropriate prototype. They even demonstrate an ability to map degraded inputs onto the appropriate prototype. This is possible because the prototypes have been strongly ingrained in the network by the training regime. Inputs that constitute one of the training categories are grouped according to the common statistical tendencies amongst them. Even when only a few of these are present in a degraded input there are enough to push the hidden layer activity towards one of its prototypes. In this way they can be thought of as basins in vector space, any activation vector which comes close enough rolls down into the centre, firing the prototype. The trick is to get only the right inputs falling under the influence on an attractor basin, and it is a trick at which networks seem to be remarkably adept.

However, this is not the full extent of the representational power of hidden layers. For they also exhibit a semantic metric, which is to say that the prototypes are not randomly distributed in the vector space. As an example consider a network designed by Elman which had the task of discovering the lexical-category structure of a set of words. This was done by getting the network to predict the next word in a sentence.[10] To do this the network employed an architectural feature which will be prominent in what follows, namely recurrent connections.[11] Elman used the usual

---

[10] 'Representations and Structure in Connectionist Models', in G. T. Altmann, ed., *Cognitive Models of Speech Processing: Psycholinguistic and Computational Perspectives* (MIT Press, Cambridge, Mass., 1990), pp. 345-382.

[11] The first recurrent network models were created by M. I. Jordan, 'Serial Order: A Parallel Distributed Processing Approach', Report 8604, Institute for Cognitive Science (University of California, San Diego, La Jolla, 1986). The addition of recurrent connections is a step towards more brain-like simulations, and thus a step in the right direction, as argued in chapter 2.

three layer feed-forward architecture with an additional layer of context units which received connections from the hidden layer and then sent return connections back to the hidden layer. The weights of these connections were all 1, and were not altered during training. Thus the hidden layer received a context input representing its state in the previous processing cycle, i.e. a representation of the previous word, or words, in the sentence. Put simply, recurrent connections allow the network to operate in the temporal dimension.

The net was trained with crude three word sentences, presented a single word at a time, and its task was to predict the next word. The inputs to the network were localist, i.e. each word was represented by the a single unit in a 31 unit input layer, so that the network had no obvious clues as to its grammatical category or meaning. What is interesting about this network is the way it responded to the task, given a sentence such as: 'man eats. . . ,' the network activated all of the output units corresponding to words for edible things. Thus it seems that the network had learned a syntactic/semantic category. An analysis of its hidden units was done by taking the hidden layer activation patterns corresponding to each word and measuring the distance between each pattern and every other pattern. This information was then used to form a hierarchical structure, as shown in Figure 3, on page 16. Words that are close together in activation space are on adjacent branch endings, whereas words which are far apart in vector space are on different branches. This analysis revealed relations amongst word prototypes reflecting their semantic categories. At the broadest level (the first branching) there was a partition between verbs and nouns, but even beyond this the network had grouped the prototypes in highly interesting ways, into animate and inanimate nouns, and as demonstrated in the example given above, words for edible things were clustered together in the hidden layer vector space. For the purposes of this analysis an average of a word's position in vector space was used, because the positions varied with context. Far from being a drawback this variation constitutes an advantage because it allows the network to be sensitive to these contexts and respond with appropriate modifications of its output.

The upshot of all this is that the generalizations which the network made were highly structured. The spatial relations amongst prototypes in the hidden layer vector space are highly abstract, reflecting complex patterns in the input. Through the partitioning of its vector space the network acquired an ability to recognize highly

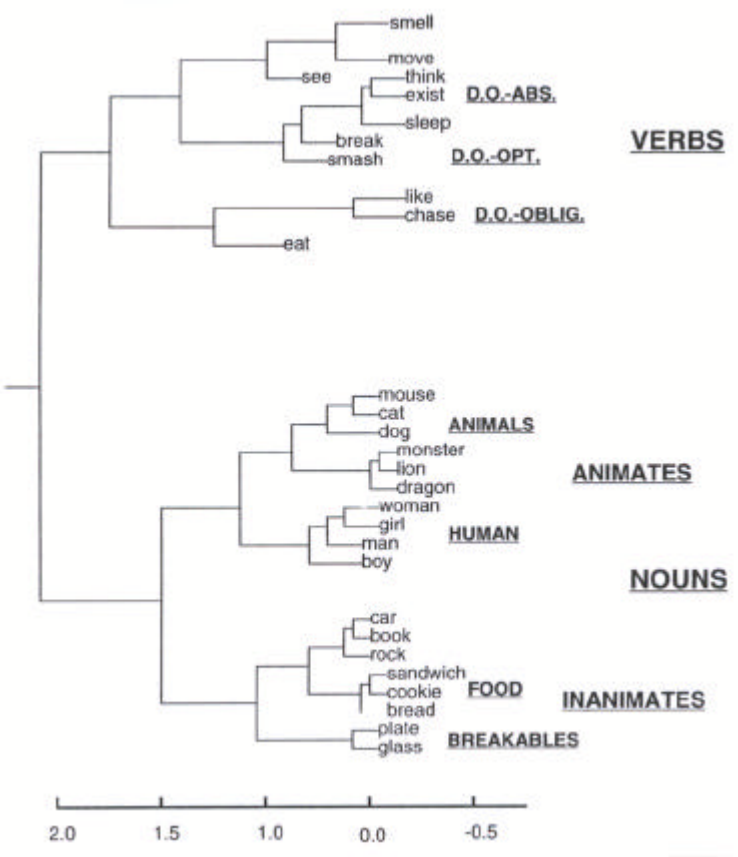

Figure 3: Hierarchical clustering of hidden unit activation patterns, demonstrating the semantic metric of the network's hidden unit vector space.

abstract features of its input. This structure in the hidden layer was then utilized by the output layer to produce the appropriate responses as defined by the assigned task. Considering Chomskian arguments about the poverty of the linguistic stimulus available to infants, this network should act as a warning that the learning environment may in actual fact carry much more information than was previously thought.

Distributed representation not only gives rise to complex processing abilities, it also means that networks are operationally robust. As representation and computation involve many units and connections, each individual unit or connection plays only a small part in the overall calculation. As a result, a network can afford to lose a few connections or units whilst still attaining a reasonable level of performance. It can also cope with incomplete or noisy inputs. The performance gradually declines with the number of elements that have been removed, or the disturbance to the input, hence the processing of distributed networks is described as displaying 'graceful degradation'. This feature is noteworthy because of its relevance

to biological systems. First real neurons are inherently 'noisy' in that they have a basal spiking frequency from which they deviate according to a normal distribution. Thus a neuron may fire more vigorously, in a way that would normally have some significance for its receiving neurons, even when none of its preferred stimuli are present. Second, it is an unfortunate fact of life that brains lose neurons throughout life, whether due to injury or the natural course of ageing, and so must be able to continue to function in the face of such loses. Put bluntly, the connectionist argues that, as brains degrade gracefully, anything claiming to model the brain had better do the same. The contrast here is supposed to be with conventional computers, which will crash even if only one line of a program, or one transistor, is missing or broken. While this is a persuasive argument it is far from conclusive, I merely mention it as an important debate between parallel distributed processing and conventional computational models.

1.2 Vector Cognition

A bare description of the functioning and abilities of real and artificial networks does not count as philosophy. To be relevant to philosophy the processing of these networks must be related to mental processes occurring at the personal level. Paul Churchland has suggested that cognition should be modelled on the way basic perceptual processing works.[12] According to his view of such processing, sensory inputs are mapped onto the appropriate prototype, which constitutes the creature's understanding of that input. Thus the bridging principle between neurocomputation and the philosophy of mind is that of the prototype. In this section I will give a brief sketch of Churchland's approach before going on to show that it, and other standard connectionist models, cannot fully account for human cognition in section 1.3.

1.2.1 Vector Coding

The first point that Churchland makes is about the power of vector coding. He illustrates this by considering the way in which the brain represents colour. The human retina contains three kinds of colour sensitive cells, each one being maximally

[12] *A Neurocomputational Perspective: The Nature of Mind and the Structure of Science* (MIT Press, Cambridge, Mass., 1989), Ch. 10.

responsive to a different wavelength of light. These cells project to a population of neurons which contain three further types of cells, and the connections are such that they demonstrate opponent processing. This means that the activity of the cells signals the presence of either of two colours, but not both simultaneously. For example, one type is red-green opponent, which means that they increase their spiking frequency when stimulated by red light and decrease their spiking frequency when they are stimulated by green light. The second type of opponent process cells is yellow-blue opponent, and the third register the relative brightness values across all three colour photoreceptor cell types. The upshot of all this is that each of the types of opponent process cells can be considered to form an axis of human colour space. Thus any colour that can be perceived by humans is represented by a point in that three-dimensional space determined by the relative activities in the three cell types.

Colour space has only three dimensions, but if we make a conservative assumption that there are only 10 distinct positions on each axis this gives 1000 different positions in the vector space.[13] Thus the representational power of even a small group of neurons is staggeringly huge, and rises exponentially. Indeed, humans have four types of taste receptor cells, and at least six types of olfactory receptors, which goes some way to explaining the sensitivity and sophistication of our sensory capacities.

Vector coding need not be limited to the representation of such basic sensory features. With the use of many more dimensions highly complex domains can be represented. For example, there is a reasonable body of evidence to suggest that there is a specific region of the brain involved in the representation of faces.[14] Thus this area might constitute a face vector space with each point representing a particular face.

A caveat is required here concerning the explanatory power of vector coding, contra Churchland, because to say that a sensory experience just is a certain pattern of spiking frequencies in a population of neurons is not fully to explain what is a

---

[13] Evidence from psychophysical studies suggests that we can distinguish at least 10,000 colours, which suggests that there should be approximately 20 positions on each axis.

[14] See, for example, G. G. Baylis, E. T. Rolls, and C. M. Leonard, 'Selectivity Between Faces in the Responses of a Population of Neurons in the Cortex in the Superior Temporal Sulcus of the Monkey', *Brain Research* 342, 91-102.

conscious phenomenon.[15] I do not claim that the theory expounded here closes the explanatory gap between scientific description and subjective experience. For my purposes we must simply take it for granted that the activation pattern in a given population of neurons produces a subjective experience. The power of the approach lies in its capacity to make this leap of faith seem less daunting, and ultimately plausible. The key to the representational power of such neuronal activation vectors lies in the way in which they can be transformed from one population to another via the complex connections between them in a way that respects their informational capacity and content, i.e., not according to some abstract syntactic aspect of the representation, but through vector transformation, via a tensor matrix from one neuronal activation vector to another.

1.2.2 Vector Prototypes and Explanatory Understanding

Churchland champions the idea that explanatory understanding 'consists in the activation of a specific prototype vector in a well-trained network.'[16] That the network be well-trained is crucial, for it explains why the activation of the prototype is not merely a labelling of the input stimulus. The activation of a prototype:

> . . . represents a major and speculative *gain* in information, since the portrait it embodies typically goes far beyond the local and perspectivally limited information that may activate it on any given occasion. That is why the process is useful: it is quite dramatically ampliative. On each occasion, the creature ends up understanding (or perhaps *mis*understanding) far more about the explanandum situation than was strictly presented in the explanandum itself. What makes this welcome talent of ampliative recognition possible is the many and various examples the creature has already encountered, and its successful generation of a unified prototype representation of them during the course of training.[17]

This allusion to the role of previous encounters has an interesting echo in Evans' notion of an informational system, which he saw as central to explaining the nature of thought about particulars.[18] Evans rightly stresses the relationship between information and recognitional capacities:

---

[15] This argument is made by P. M. Churchland in 'Reduction, Qualia, and the Direct Introspection of Brain States', *Journal of Philosophy* 82 (1985), 8-28.

[16] *A Neurocomputational Perspective*, p.210.

[17] *A Neurocomputational Perspective*, p. 212.

[18] *The Varieties of Reference* (Oxford University Press, Oxford, 1982), Ch. 8.

. . . we should expect that, in any system in which information is stored about particular objects, there will be a central core of cases in which the subject has associated information with a capacity to recognize a particular individual . . . These are the paradigm cases: evolving clusters of information generated in a pattern of encounters in which the recognitional capacity was triggered, and still linked with that capacity, which serves as the means to identify opportunities for using old, and gaining new, information.[19]

In Churchland's model we have an obvious explanation of how this informational system operates, and what is more, operates rapidly and robustly.[20] The recognitional capacity and its associated information can be united as a vector prototype. Encounters with an object (as an individual, or as a token of a type or category) bring about activation of the prototype, and each encounter provides an opportunity for changes to be made in the configuration of the hidden layer vector space, to encode any novel information. In this way the prototype could come to include various kinds of expectations that go beyond the present experience, and play a role in the control of on-going behaviour, as Churchland describes:

The picture I am trying to evoke, of the cognitive lives of simple creatures, ascribes to them an organized 'library' of internal representations of various prototypical perceptual situations, situations to which prototypical *behaviours* are the computed output of the well-trained network. The prototypical situations include feeding opportunities, grooming demands, territorial defence, predator avoidance, mating opportunities, offspring demands, and other similarly basic situations, to each of which a certain broad class of behaviours is appropriate. And within the various generic prototype representations at the appropriate level of hidden units, there will be subdivisions into more specific subprototypes whose activation prompts highly specific versions of the generic form of behaviour . . . These various prototypes are both united and distinguished by their relative positions in the hidden-unit vector space. They are all close together, but they differ slightly in their positions along one or more of the relevant axes. These differences evoke relevantly different responses at the output layer.[21]

Thus one of the key advantages of Churchland's account is that representations are not conceived of as inner models of reality, which must then lead to action through a further distinct process of cogitation involving the model. Rather part of their identity

---

[19] *The Varieties of Reference*, pp. 276-7.

[20] There is nothing in what Evans wrote that commits him to this or any other explanation of the substrate for informational system. I draw the comparison because I think that Churchland's ideas suggest how the informational system might be implemented. I would argue that an understanding of the details of this implementation enriches and alters the account pitched at the genuinely philosophical level.

[21] *A Neurocomputational Perspective*, p. 207.

is constituted by the behaviour which they produce; representations are action-centred.[22] As a result, perception and action can be regarded as essentially cognitive processes, they do not constitute the links between the mental and the external world; they are a part of the cognitive processing itself. In the case of perception, there is no sensory Given, which must then be interpreted.[23] Incoming sensory data is processed by being mapped onto a prototype vector. This provides an explanation for why perceptual experience bears cognitive significance; we do not see colour patches, we see objects and their affordances.[24] Again, this fits nicely with Evans' notion of information-based thought, where this is defined as follows:

> a bit of information (with the content *Fx*) is in the controlling conception of a thought involving a subject's Idea of a particular object if and only if the subject's disposition to appreciate and evaluate thoughts involving this Idea as being about an *F* thing is a causal consequence of the subject's acquisition and retention of this information.[25]

The activated prototype is the information which controls the thought, because of the part it plays in the control of ongoing cognition, interacting with other internal processes to produce an appropriate response. For Churchland the role of the prototype in processing could be extremely complex, with context and goals influencing the informational flow, through the impact of recurrent pathways, which he rightly views as important. If I understand correctly an example might be as follows: if a squirrel is hungry, the activation of its acorn prototype by an appropriate visual input would initiate feeding behaviour, whereas if it is satiated it might prompt storing behaviour. This type of contextual sensitivity might be achieved via an input to the creature's hidden layer from neural areas other than those concerned with basic sensory input. Through an appropriate learning history these additional inputs could come to affect the network's processing, producing advantageous behavioural consequences.

---

[22] See A. Clark, *Being There: Putting Brain, Body, and World Together Again* (MIT Press, Cambridge, Mass., 1997).

[23] See W. Sellars, 'Empiricism and the Philosophy of Mind', in H. Feigl and M. Scriven, eds., *Minnesota Studies in the Philosophy of Science*, vol. 1 (University of Minnesota Press, Minneapolis, 1956), pp. 253-329, and J. McDowell, *Mind and World* (Harvard University Press, Cambridge, Mass., 1994) for discussion of the Given.

[24] See J. J. Gibson, *The Ecological Approach to Visual Perception* (Houghton Mifflin, Boston, 1979).

[25] *Varieties of Reference*, p. 122.

On this view, the representational power of the prototype comes from its position in a richly structured recurrent vector space. Churchland makes much of this architectural feature, no doubt inspired by its ubiquity in the brain, and by the fact that it allows complex temporal behaviours, as demonstrated by Elman's network, which was discussed in section 1.1. One feature of recurrent networks, which he notes, is their tendency to settle into a repetitive cycle, known as a limit cycle, when started off by a momentary input of the right kind. Churchland hypothesizes that this could be used for the easy production of cyclical motor activities such as walking, or the beating of the heart, because they can be represented by loops in a suitable vector space (such as a joint-angle or motor-neuron space). The possibility of iterated processing cycles through a network allows trajectories and loops to be followed through vector hyperspace, reflecting the continuous cyclical nature of these activities. Recurrent connections will prove to be important in the account of cognition which follows, for as Churchland points out, it is highly likely that they have applications outside the sphere of motor control. Elman's recurrent word prediction network, which was described in section 1.1 is an example of how Churchland sees recurrent connections underlying complex cognition. This approach seems to imply a kind of representational holism, in that an activated vector is a point on a trajectory in vector hyperspace. Thought, which is temporally extended, consists in such trajectories, which are produced by recurrent cycles through an immensely complex hidden layer. Depending on context the processing will lead off in a variety of directions, and it is this diversity of trajectories, and their semantically relevant geometrical relations that are supposed to make this process so powerful. Once you have an internal array of recurrent connections which can provide their own input, and thus which can function without external input, you have the possibility for complex and abstract cognition which is not tied to current external stimuli. There need be no simple progression from input to processing to output, instead the output to any given processing cycle could be the initiation of further processing. Churchland places a great deal of stress on the effect of recurrent connections, and has suggested that such a system of recurrent networks might be the basis for consciousness for this, amongst other, reasons.[26] I do not want to take any firm

[26] *The Engine of Reason, the Seat of the Soul* (MIT Press, Cambridge, Mass., 1996), Ch. 8.

position on this claim, as I am not attempting to argue for an explanation of consciousness. However, it is clear that a flexibility in cognitive ability does seem to be a feature of conscious creatures, and so it is genuinely interesting, and suggestive, that recurrent networks seem to have a variety of interesting capabilities.[27]

One of these is the capacity to interpret the same sensory input in a number of different ways. The human capacity for this is famously demonstrated by ambiguous pictures such as the duck/rabbit. A non-recurrent network will always respond to a stimulus in the same way, but Churchland argues that a recurrent network can echo this human ability because the recurrent connections provide a means of modulating processing, either by providing a duck context or a rabbit context. Churchland goes on to argue that this capacity of recurrent networks is not only important in the context of perception. He claims that it also has impact in much more abstract aspects of cognition, in that it offers an explanation for the processes of rapid understanding and reconceptualisation. This problem has a fine tradition in philosophy, for Wittgenstein wrestled with these phenomena in the form of understanding in a flash:

> 'What happens when a man suddenly understands?'—The question is badly framed. If it is a question about the meaning of the expression 'sudden understanding', the answer is not to point to a process that we give this name to.—The question might mean: what are the tokens of sudden understanding; what are its characteristic psychical accompaniments?[28]

Normal network learning takes thousands of cycles, and can be understood as a sort of gradient descent process, as described in section 1.1. This clearly cannot explain how an individual can, literally in an instant, see a recalcitrant problem in an entirely new and fruitful way. Churchland suggests a solution to this problem based upon the idea of recurrent connections. He illustrates this with Huygen's realisation that light can be understood as a wave phenomenon:

> Here the theory of waves in mechanical media — a theory already well-formed in Huygen's mind in connection with water waves and sound waves — was applied in a domain hitherto unaddressed by that framework, and with systematic success. There was no need for Huygens to effect a global

---

[27] This would be compatible with Dennett's Multiple Drafts model, as expounded in *Consciousness Explained* (Little Brown, Boston, 1991). If it is reduced to a model of cognitive processing (access consciousness) rather than of phenomenal consciousness there seems no reason why a system of recurrent networks could not instantiate such a parallel, multitrack system. Whether a recurrent system can be the basis for phenomenal consciousness, in humans at any rate, is an interesting, but moot point for my purposes.

[28] *Philosophical Investigations* § 321.

reconfiguration of his synaptic weights to achieve this conceptual shift. He had only to apprehend a familiar class of phenomena in a new cognitive context, one supplied largely by himself, in order to have old and familiar input-unit vectors (those concerning light) activate hidden-unit vectors in an area of his conceptual space quite different from areas they had previously activated. The difference lay in the context-fixers brought to the problem.[29]

Churchland is somewhat vague about the way in which this conceptual redeployment occurs. Given that his explanations are couched in terms of recurrent hidden layer architectures, it is not clear how a model could be created which would search through its prototypes in quite this way. It is at this point that problems begin to appear for Churchland's account, although the problem of explaining the way that human cognition manages to focus down onto only the relevant options is not restricted to connectionist models. One way to see the force of the worry here is in the context of an apparent advantage of the vector prototype model: it does not present any fundamental bifurcation in nature between humans and other animals. We can make sense of the behaviour of non-linguistic creatures in terms of vectorial prototypes, as made clear in the passage quoted above. Thus we can treat (some) creatures as having genuinely representational cognitive processes without the need to attribute symbolic abilities to them. And of course the vector prototype model is ideally suited for explaining how behaviour is smoothly and efficiently carried out.

Trouble arises for this harmonious picture in that there *is* an essential difference between animal and human thought: humans use a type of representation that is unique, namely symbolic representation. In the next section I will argue that Churchland's vector prototype model cannot successfully explain some aspects of symbolic representation, such as compositionality. This becomes clear when examples such as the one quoted above involving Huygens are examined carefully. What Churchland is trying to explain here is conceptual redeployment, but to have concepts one must have compositionality, amongst other features, and for reasons which will be given in the next section, the vector prototype model cannot accommodate them. However, this shouldn't lead to a total rejection of the model, and in what follows I will attempt to show how the problems raised in section 1.3 can be overcome by emendations inspired by the processing of real neural networks.

---

[29] P. M. Churchland, 'Learning and Conceptual Change', in A. Clark and P. Millican, eds., *Connectionism, Concepts, and Folk Psychology* (Clarendon Press, Oxford, 1996), p.23.

1.3 Connectionism and Systematicity

The main accusation that is made against connectionist models is that they cannot explain the essentially structured nature of thought. That thought involving concepts must be essentially structured is a premise which I will treat as immutable. It is something that the majority of philosophers agree upon, and as such must be explained by any aspiring theory of cognition. The articulation essential to conceptual thought is expressed by Evans' generality constraint: to be ascribed the thought that '*a* is *F*' one must also be able to entertain other thoughts involving '*a*', such as '*a* is *G*', and other thoughts involving '*F*', such as '*b* is *F*'.[30] For those attempting to model thought this gives rise to what Fodor and McLaughlin call the systematicity problem:

> The systematicity problem is that cognitive capacities come in clumps. For example, it appears that there are families of semantically related mental states such that, as a matter of psychological law, an organism is able to be in one of the states belonging to the family only if it is able to be in many others. Thus, you don't find organisms that can learn to prefer the green triangle to the red square but can't learn to prefer the red triangle to the green square. You don't find organisms that can think the thought that the girl loves John but can't think the thought that John loves the girl. You don't find organisms that can infer *P* from *P&Q&R* but can't infer *P* from *P&Q*. And so on over a very wide range of cases.[31]

There is a considerable amount of literature devoted to the subject of systematicity, and it involves several strands.[32] One concerns the exact nature of systematicity, and another whether connectionist architectures can be genuinely systematic. Difficulties arise in the context of the first of these because one cannot replace a verb's argument with any old word, there are various kinds of constraints, and these must be learned.[33] This suggests that systematicity should admit of degrees, but this does not appear compatible with Fodor and McLaughlins' position. Suffice it to say, for present

---

[30] *Varieties of Reference* (1982), Ch. 4.

[31] 'Connectionism and the Problem of Systematicity: Why Smolensky's Solution Doesn't Work', in C. Macdonald and G. Macdonald, eds., *Connectionism: Debates on Psychological Explanation* (Blackwell, Oxford, 1995), p. 200.

[32] See, for example, J. Fodor and Z. Pylyshyn, 'Connectionism and Cognitive Architecture: A Critical Analysis', in S. Pinker and M. Jacques, eds., *Connections and Symbols* (MIT Press, Cambridge, Mass., 1988), pp. 3-71, A. Clark, 'Systematicity, Structured Representations and Cognitive Architecture: A Reply to Fodor and Pylyshyn, in T. Horgan and J. Tienson, eds., *Connectionism and the Philosophy of Mind* (Kluwer Academic Press, Boston, 1991), pp. 198-218, and R. F. Hadley, 'Systematicity in Connectionist Language Learning', *Mind and Language* 9 (1994), 247-272.

[33] See S. Pinker, *Learnability and Cognition: the Acquisition of Argument Structure* (MIT Press, Cambridge, Mass., 1989).

purposes, that the feature of thought that I am interested in concerns some kind of sensitivity to structure, and an ability to apply information in relevant, but different, task domains. I hope that the following discussion will give an intuitive feel for the problem at hand, even if it does not furnish a precise definition.

This type of systematicity cannot be satisfactorily accounted for by the processes of generalization, pattern completion, and prototype extraction which form the basis of Churchland's approach. A network may have prototypes for 'red' and 'square' but how does it represent the complex thought that the square is red? This could not be done by having both prototypes activated at the same time because their vectors are fully distributed, and so each one involves the activity of all the units. The two patterns cannot just be added together, because that would merely produce a vector between the two original ones. This intermediate vector would have a representational significance determined by the semantic metric of the vector space, and given the hyper-dimensional nature of the metric, it would be a serendipitous accident if this happened to be the representation for 'red square'. The hyper-dimensionality rules out vector addition because the semantic significance of any point in vector space is determined by its relations to all of the dimensions of the space, not just a few of them. Presumably an additional 'red square' prototype would be needed, but even granted the enormous representational power of hyper-dimensional vector spaces this multiplication of prototypes could not proceed indefinitely. Even if it could this would be no solution to the problem of systematicity, for it would mean that every complex thought would be represented by an unarticulated vector. The whole force behind the problem is that the *same elements* appear in many different complex thoughts.[34] Geach uses an analogy with chess to illustrate the relationship that must exist between thoughts and their constituent concepts:

> Making an appropriate move from a certain position may be, and at the opening of the game very likely will be, a learned response; but in the middle game it will certainly not be so, for the position may well occur only once in a life-time of play. On the other hand, the ability to make an appropriate move from a given position always presupposes a number of simpler, previously acquired, skills — the

[34] This line of argument is powerfully deployed by J. Fodor and Z. Pylyshyn in 'Connectionism and Cognitive Architecture: A Critical Analysis'.

capacities to carry out the moves and captures that are lawful for the pawns and the various pieces. As these skills are related to the chess-move, so concepts are related to the act of judgement.[35]

One might address this by resorting to localist microfeatures, so that there would be microfeatures for redness and squareness, and they would be co-active when representing a red square. As it stands this is no solution, because the lapse into the use of localist microfeatures as the dimensions of the representational vector space is self-defeating because the semantic contents of those dimensions are left unexplained.[36] A red microfeature might gain representational significance through causal correlation, but this does not provide adequate resources for explaining how the semantic content of complex relations such as 'loves' could be represented as a microfeature. I will not rehearse the inadequacies of causal theories of meaning here, for even if the point is granted a second problem remains: how can a set of co-active microfeatures ('John', 'loves', and 'girl') represent the thought that John loves the girl rather than that the girl loves John? It would seem that a 'John-as-subject' microfeature is needed, but here again we are faced with an explosion of representational elements, one for each possible syntactical position. Given that language is productive, having no definite upper boundary, this seems an implausible explanation.

An attempt to rescue the original line of argument might be made by maintaining that the various prototypes are not as unstructured as they first appear. One might imagine a 'John' region of the vector space, with different grammatical relations being represented by different positions within this region. After all, analysis of Elman's word prediction network, discussed in section 1.1, showed that its hidden layer was highly structured, with major divisions and subdivisions reflecting lexical-category structure. Thus the articulation of conceptual thought would be captured by the semantic metric and the relations of a point to many others, and through the effect of further content on the trajectory plotted through vector space. However, doubt is cast on this suggestion by the inability of networks to apply knowledge gained in one situation to good effect in a different situation. For example, Elman's network

---

[35] *Mental Acts: Their Content and their Object* (Routledge and Kegan Paul, London, 1957), p.13.

[36] I am assuming that the semantic metric of distributed representations account for their semantic content. This is a bold claim that requires argument, and this will come later in section 3.3. For the moment I want to put the issue to one side whilst dealing with the inadequacies of a localist approach to the problem of meaning.

structured its hidden layer vector space into nouns and verbs, but there is no easy way to utilize this if we wanted to use the network to categorize a set of words into verbs and nouns. All it can do is 'predict' what type of word will come next in a sentence; the network's expertise is extremely domain specific. Clark relates this to the notion of non-conceptual content, as developed by Cussins.[37] Non-conceptual content is defined by Cussins as content which consists of non-conceptual properties, where conceptual and non-conceptual properties are defined as follows:

> A property is a conceptual property if, and only if, it is canonically characterized, relative to a theory, only by means of concepts which are such that an organism *must have* those concepts in order to satisfy the property.
>
> A property is a non-conceptual property if, and only if, it is canonically characterized, relative to a theory, by means of concepts which are such that an organism *need not have* those concepts in order to satisfy the property.[38]

These notions can be used to express the problem with Elman's network. Whilst the relations revealed in its hidden layer space can be described in terms of the concepts of 'noun' and 'verb' they only have non-conceptual content. Cussins also provides the resources to argue for this claim. Progression from non-conceptual to conceptual content is marked by increasing perspective independence, a point illustrated by consideration of frogs and their fly-detecting abilities:

> The frogs' 'fly-thoughts' are not really fly thoughts because their success (and hence their content) depends on special features of the frog task-domain (the cost of tongue-swipes at massive distant objects is outweighed by the benefit of successful fly catches); frog 'cognition' is dependent on the perspective of a particular task-domain. It cannot generalize.[39]

If the frogs were placed in a different task-domain, where tongue-swipes are more costly, perhaps attracting predators, they would soon be in trouble. They cannot alter

---

[37] *Associative Engines*, p. 73.

[38] 'The Connectionist Construction of Concepts', in M. Boden, ed., *The Philosophy of Artificial Intelligence* (Oxford University Press, Oxford, 1990), pp. 382-3. Although this notion is similar to the one developed by C. Peacocke in *A Study of Concepts* (MIT Press, Cambridge, Mass., 1992), there is an important distinction that has been debated in the literature, about whether a creature can posses non-conceptual content if it has no conceptual states at all. This is something that is denied by Peacocke, but acceptable on Cussins definition as stated above; see J. L. Bermúdez, 'Peacocke's Argument Against the Autonomy of Nonconceptual Representational Content', *Mind and Language* (1994), 402-418, for a discussion of the issue. I side with Cussins on this debate, as the prime reason for introducing non-conceptual content is to explain how creatures can have genuinely representational states without having to meet the generality constraint.

[39] 'The Connectionist Construction of Concepts', p. 424.

their behaviour in response to environmental changes. If, on the other hand, frog fly-detectors were sensitive to many other abilities, and to environmental circumstances, then they would be on the way to achieving conceptual content. Thus perspective independence can be seen as the capacity to apply one's abilities and knowledge in a way that allows one to cope with disturbances in the task domain. One can bring these abilities to bear from any angle, and in any place. I do not think it I overly contentious to hold that this cognitive robustness is at the heart of what makes humans intelligent creatures; it accounts for the evolutionary advantage of being intelligent, because it allows us to survive in an ever changing environment.

The notion of perspective independence can be applied in exactly the same way to connectionist networks: they are extremely dependent on their task-domains — slight changes completely destroy their ability to respond appropriately. Elman's network cannot be used in any situation other than the one in which it was trained, and so it cannot meet the generality constraint, and so cannot be considered to have the concepts of 'noun' and 'verb'. Perspective independence as a criterion for conceptual content is really just another way of expressing the generality constraint. To count as conceptual a thought must contain elements that can be combined with many other such elements in ways that respect their semantic values. This gives us both a handle on the problem with connectionist models and a criterion for establishing when it has been overcome. This is useful, because despite the difficulties raised in this section for Churchland's model, it would be wrong to assume that they rule out any parallel distributed processing account. I will suggest an alternative approach in chapter 3.

1.4 Summary

In this chapter I have explained how connectionist networks operate, and how Churchland has extrapolated from this to a model of cognition. I then pointed out a few problems for this model in explaining the systematicity of human thought. The following points will be important in what follows:

- The distributed nature of vector representation, which is the source of many of the abilities of connectionist networks.
- The encoding of experience by the weights of connections.

- The notion of a vector space and its incorporation of a semantic metric, with vector prototypes arranged in that space in a way that reflects their semantic relations.
- The distinction between conceptual and non-conceptual content and its relation to the systematic nature of human thought.

2 Biologically Plausible Neural Computation

The discussion in the previous chapter dealt with mainstream connectionist networks. These provide valuable insights into some aspects of cognition, but for other aspects, they may even prove misleading. The problems with such networks arise because of their lack of biological plausibility. The main reason why it is unlikely that these sorts of networks are implemented in the brain is their use of non-local learning algorithms, most commonly the backpropagation algorithm. These algorithms are non-local because, as described in section 1.1, they compare the network's actual output with the correct output. This information about the correct result is not available locally in the network, and must be provided by an external teacher. The error signal is propagated back to connections in earlier layers of the network, further compounding the non-local nature of the learning mechanism. Given current knowledge of neuronal architecture in the brain there is no way that this sort of precise information could be provided.

The neuronal plausibility of connectionist models is further strained by the fact that single neurons cannot produce both excitatory and inhibitory connections, because all of their synapses must use the same neurotransmitter. Thus neurons are either excitatory or inhibitory, but not both. The vast majority of connectionist models allow connections from a single unit to take both positive and negative weight values. Most connectionists would not see these architectural divergences as a problem. Their standard response is to argue that they are modelling higher level cognitive processing, and that this necessitates only a loose connection between models and reality. Marr expressed this point in the following way:

> Trying to understand perception by studying only neurons is like trying to understand bird flight by studying only feathers: it cannot be done. In order to understand bird flight, we have to understand aerodynamics; only then do the structure of feathers and the different shapes of birds' wings make sense.[40]

---

[40] *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information* (W. H. Freeman, San Francisco, 1982), p. 27.

I agree with this viewpoint, as I made clear in the introduction; the whole purpose of modelling is to strip out the unnecessary details, so that the global picture can be seen more clearly. Thus connectionist models should not be slavish duplications of real neural networks. My objection, which is central to my thesis, is that some features of real neural networks that have been abandoned suggest interesting new ways of understanding the emergent behaviour of networks. It is precisely by paying attention to the details that we get a better grip on the large-scale principles at work in real brains, and that produce real thinking.

Instead of using sophisticated non-local learning algorithms, recent research into neural networks has suggested that they overcome computational problems through the use of a number of different architectures linked together, with a Hebbian learning rule, and sparse coding (see section 1.1). The Hebb rule states that where a presynaptic neuron and a postsynaptic neuron are active at the same time, the synapse that connects them will be strengthened according the following algorithm:

$$\delta w_{ij} = k\, r_i\, r¢_j.$$

where r_i is the postsynaptic firing rate, $r¢_j$ is the presynaptic firing rate, k is a learning rate constant, and δw_{ij} is the change in the synaptic weight w_{ij} (this nomenclature, in which the ith neuron is the postsynaptic neuron, and the jth neuron is the presynaptic neuron, is standard). For Hebbian learning all the information required is available locally at the synapse, and empirical research has revealed a plausible mechanism at the microcellular level for this rule. At least three different neural architectures have been found in the brain which utilize this basic learning rule to accomplish different sorts of computational tasks: pattern association networks, autoassociation networks, and competitive networks, examples of which are shown in Figure 4, on page 33. These architectures are all compatible with connectionist modelling; it is the use of non-local learning algorithms that is responsible for the lack of biological plausibility. The divergence from connectionist networks arises when these architectures are combined with a Hebbian learning algorithm.[41]

I will briefly describe biologically plausible pattern association and autoassociation networks in sections 2.1 and 2.2 respectively, before going on to

[41] It is true that some models have used a Hebbian learning algorithm, but not in the ways that will be described below.

discuss competitive networks in more detail in section 2.3, because they play a more central role in complex cognitive processing. Then, in section 2.4, I will go on to discuss the interactions of these different architectures in actual brain systems.

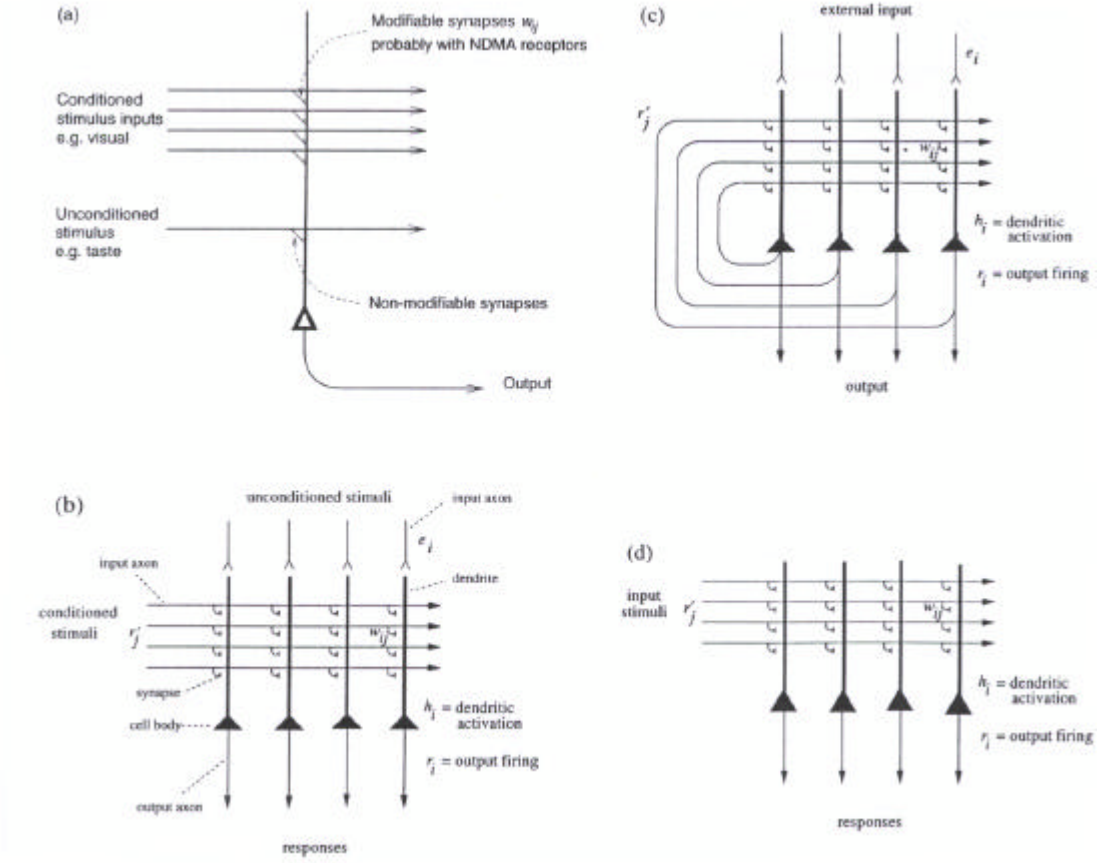

Figure 4: Examples of three network architectures that use local learning rules: (a) Pattern association with a single output neuron; (b) Pattern association network; (c) Autoassociation network; (d) Competitive Network.

2.1 Pattern Association Networks

Pattern association networks receive two sets of synaptic input: an unconditioned and a conditioned input. The former has unmodifiable synapses, as its name suggests, and represents a primitive reward signal which is 'hard-wired' into the brain. This input is effectively dominant, when active it determines the pattern of activity on the output neurons. The conditioned input has modifiable synapses. In terms of the connectionist networks described in chapter 1, these can be thought of as two different sets of units with connections onto a single set, or layer, of units.

These pattern association networks are believed to underlie conditioned learning. Hence the task of such networks is to associate a vector on the conditioned inputs with the same output vector as is produced by an unconditioned input. For example, the taste of food, which is intrinsically rewarding might be paired with the sight of the food so that the sight of the food also comes to be rewarding. The unconditioned stimulus acts in some way as the teaching input, forcing the desired output vector. As described in section 1.1, neurons can be viewed as comparing input vectors with their weight vectors, firing if they are sufficiently similar. Through Hebbian modification the weight vectors of the network's neurons come to reflect the input vector which was co-active with the unconditioned input vector, so that the conditioned input vector comes to elicit the same response.

2.2 Autoassociation Networks

As can be seen in Figure 4, on page 33, the distinguishing feature of autoassociation networks is the fact that each neuron is connected to every other neuron in the network. The task of these networks is to produce an output firing vector which is the same as their external input vector. This might seem pointless until it is noted that these networks have an ability to produce a complete output vector from only a fragment of a previously presented input vector. Thus autoassociation networks are believed to form the basis of both episodic and short term memory.

Autoassociation networks work by storing associations between the elements in an input pattern. The addition of recurrent connections makes such systems dynamic in nature, and their operation can be understood in terms of attractor basins similar to those which were appealed to in section 1.1 to explain the prototype extraction of connectionist networks. The similarity is not total, however, because of the dynamic nature of autoassociation networks. In feedforward connectionist networks, the units in the hidden layer are considered as attractors because they draw together patterns on the input layer. Thus such attractors emerge over time through training, but on any one trial presentation they operate over a single time cycle. By contrast, in autoassociation networks the movement towards the bottom of a basin occurs as patterns of activation cycle round the system over many iterations until a stable state is reached, rather like a ball rolling down into a valley.

This dynamic behaviour can be understood by considering pairs of neurons. If they are both firing above their base rate, and they have a positive connection weight, then they will reinforce each other and contribute to the stability of the system. If one were firing below its base rate, or the connection had a negative weight, then they would be unstable and would tend to 'struggle' to achieve dominance, trying to switch each other's activation state to support their own. These sorts of processes would occur throughout the network, until the least antagonistic state is reached, i.e. the state with the most compatible relationships between neuron pairs. Through Hebbian learning the weights are altered so that these states, the bottoms of the attractor basins, come to reflect those input patterns that were presented to the network.

The capacity of autoassociation networks (in terms of patterns stored without significant interference) is a complex matter, but Rolls and Treves have shown that for biologically plausible networks it is a function of the number of recurrent connections, and the sparseness of the patterns, according to the following equation:

$$p \approx \frac{C^{RC}}{a \ln(1/a)} k$$

where p is the number of patterns, C^{RC} is the number of synapses onto each neuron, and a is a measure of the sparseness of the patterns, n is the number of neurons, and k is a complex factor dependent on several aspects of the network.[42] The result is that the greater the number of recurrent connections onto each neuron the larger the number of different patterns that can be stored. For example, for C^{RC} = 12 000 and a = 0.02, p is calculated to be approximately 36 000. The implications of the sparseness of representations is discussed in section 2.4.2.

2.3 Competitive Networks and Convergent Architectures

An example of a competitive network is shown in Figure 4, on page 33. Competitive networks are so-called because they utilize mutual inhibition between their output neurons. In the most extreme case this might result in only one neuron remaining active, a winner-takes-all scenario. Competitive inhibition forms the basis for a

[42] 'What Determines the Capacity of Autoassociative Memories in the Brain?', *Network* 2 (1991), 371-397.

number of features which make this architecture useful in perceptual systems. The most important of these features is self-organisation, which allows the emergence of feature detectors.

In some sense the competition takes the place of the teaching signal in connectionist backpropagation networks, in that it allows the network to extract prototypes from its input stimuli. It is easiest to explain why this is the case using the example of a winner-takes-all network. Initially such a network will have random weights on its synaptic connections. When a particular input is presented it might happen to create an activation pattern on the inputs which is closer to the weight vector of one of the neurons than to those of the others. Over a short period of time the effect of the mutual inhibition will leave only this neuron firing. On a larger time-scale, and many stimulus presentations, provided that there are genuine clusters of patterns in the inputs, the weight vectors of the neurons will come to adopt the central tendencies of those clusters, i.e. they will come to represent prototypes. This occurs through the influence of superpositional storage of the distributed input patterns, as described in section 1.1. Those synapses most often activated on the winning neuron by a cluster will become the strongest through Hebbian modification.

This competitive prototype extraction is the basis for a number of interesting properties. These can be demonstrated most graphically in the visual system, and in the pathway devoted to object recognition in particular (see Figure 5, on page 37). Recordings from single neurons within this pathway, along with information about neural architecture and connectivity, have lead Rolls to suggest a computational model of how object recognition is achieved.[43] Rolls has proposed that this is achieved by a system of hierarchically connected competitive networks. The hierarchical organisation is crucial because it allows each successive layer to extract more abstract features from increasingly large receptive fields, until cells are reached in the anterior inferotemporal cortex which respond to objects regardless of position in the visual field and viewing angle. Hence the object recognition achieved by this pathway is described as invariant.

---

[43] 'Brain Mechanisms for Invariant Visual Recognition and Learning', *Behavioural Processes* 33 (1994), 113-138.

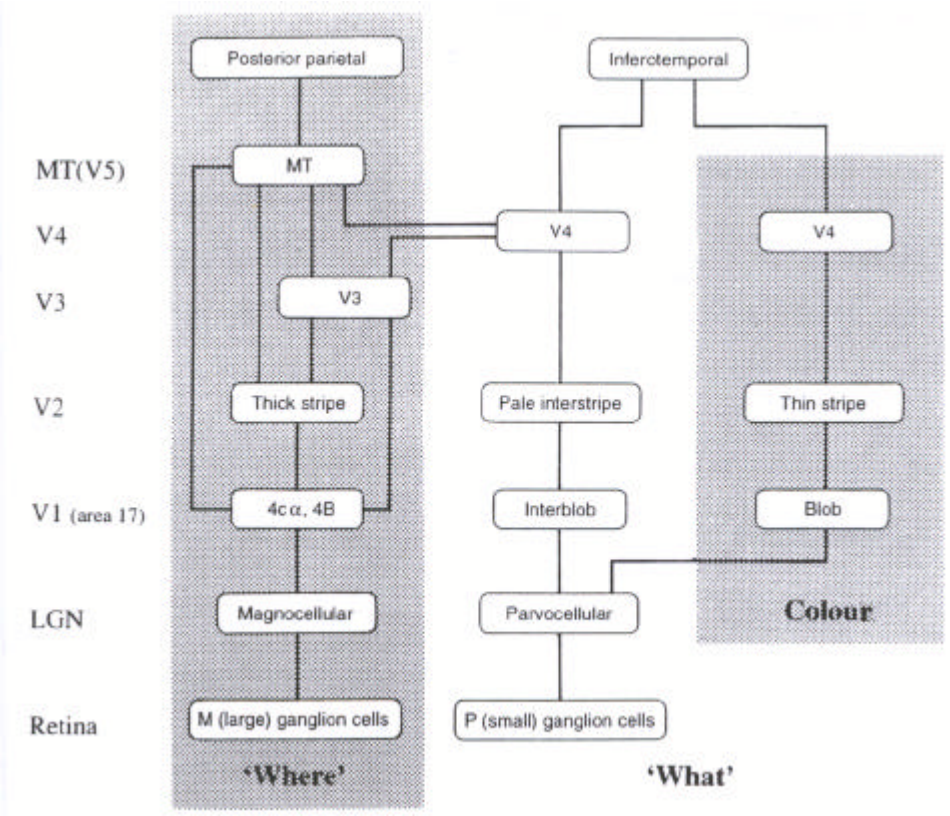

Figure 5: A schematic diagram of the visual pathways from the retina to the visual cortical areas. V1, primary visual cortex; V2, V3, V4 etc., other cortical areas; M, magnocellular, P, parvocellular.

The first evidence for this hierarchical hypothesis can be found in the early stages of the visual system. Some cells in the lateral geniculate nucleus, which is the way station between the retina and the visual cortex, have concentric on-centre off-surround (or vice versa) receptive fields, as do the ganglion cells in the retina. This means that they are excited by light falling in the centre of their receptive fields, and inhibited by light falling in the region around this centre. Some cells in area V1 have receptive fields which are not circular, but elongated, so that they are sensitive to lines or edges at certain orientations. When these patterns of receptivity were first discovered Hubel and Wiesel suggested that they could be accounted for by a convergence of connections from several cells in the lateral geniculate nucleus onto a cell in V1.[44] For a line passing over the retina at a given angle will cause adjacent retinal ganglion cells, and hence lateral geniculate cells, in that orientation to fire

[44] 'Receptive Fields, Binocular Interaction, and Functional Architecture in the Cat's Visual Cortex', *Journal of Physiology* 160 (1962), 106-154.

simultaneously. The problem for the brain is how to orchestrate the connectivity between layers to create these receptive fields.

It has been shown using simulations that this organisation can emerge with competitive network architectures. A detailed and biologically realistic of model, called VisNet, has been produced by Rolls. VisNet is effectively a reproduction of the object recognition pathway, based on current neurophysiological evidence.[45] The model consists of four layers roughly corresponding to V2, V4, the posterior temporal cortex, and the anterior temporal cortex. Each layer consists of 32×32 cells. The connections to a cell from those in the preceding layer arise from a topologically corresponding region in that layer. The connectivity decreases with distance from the centre of the receptive field. Lateral inhibition within each layer has a radius of effect just greater than the radius of the neurons' receptive fields, mirroring the pattern of connectivity of inhibitory interneurons in the visual cortex. It is these inhibitory connections which allow competitive learning. The inhibition in VisNet is set so that the competition is 'soft', meaning that the neurons produce a graded response rather than a sole winner.

Adjacent cells within a layer have overlapping receptive fields, and so their activity will tend to be highly correlated, whilst cells further away will not have highly correlated activity, because there will be no receptive field overlap. This explains how centre-surround receptive fields might emerge at the next layer of such a convergent architecture, as the neurons in that higher level detect and categorize these correlations amongst their input neurons.

The activity from the cells in this layer feeds on to those in the next, and the same pattern of correlation will be repeated, but this time the neurons are receiving from centre surround neurons, and thus they will begin to extract higher order features such as bars or edges, just as Hubel and Wiesel hypothesized. VisNet uses a version of the Hebbian learning rule, with a small but important modification, namely a short memory trace in the postsynaptic neurons. This matches the physiology of real synapses, and is thought to be useful because it allows higher level neurons to recognize correlations in the activity of their inputs caused by objects moving over

---

[45] 'Brain Mechanisms for Invariant Visual Recognition and Learning'.

the retina. For example, a line moving across the retina will cause a succession of edge-detecting neurons to fire. If the postsynaptic neuron remains active for a short period this will allow it to pick up on these patterns and thus to recognize the same feature at various different parts of the retina.

In one testing regime a set of three non-orthogonal stimuli were used, such as 'T', 'L', and '+' shapes. These shapes are non-orthogonal because they use the same elements in different combinations, thus the system was being tested to see if it had the ability to recognize spatial combinations of features, not just the bare presence of those features. VisNet was trained by sweeping the stimuli across its 'retina' in a random sequence. After learning the network was tested by examining the response profile of its neurons. Cells were found in layer 4 which responded to the presence of particular test stimuli regardless of position on the retina. VisNet was even capable of producing cells in layer 4 which would respond to the presence of a particular face, regardless of position, and which of seven possible views were presented.

This demonstrates that an architecture like the one described is capable of invariant object recognition. It matches the data from single neurons in the primate cortex, where cells have progressively larger receptive fields with each layer, until neurons in the inferotemporal cortex are reached which respond to specific objects, or types of objects, anywhere in the visual field, just like those in layer 4 of VisNet.

The product of this invariant object recognition system is a pattern of firing of a relatively small number of neurons in a moderately large population, where that population is at the top of a set of hierarchically arranged cell populations. From what has been discussed so far a number of computational advantages can be seen to follow. First, as sets of correlated inputs get represented by the activity of a few neurons in the next layer this removes redundancy in the sensory input. What required a large number of neurons to code gets represented in 'short hand' by a much reduced number of neurons. In a convergent network each neuron only has to sample a small part of the preceding layer. This avoids a combinatorial explosion in connections, which would be produced if each neuron had to search the whole of the preceding layer for the presence of its preferred feature combination. This localization of feature search also solves another computational problem faced by other models of visual recognition, namely feature binding. This is illustrated by the fact that a list of the features present in the visual scene is insufficient for object recognition, as it might be

possible for the same group of features to be differently arranged to form different objects. One must also have a way of representing the relations between features. This is achieved automatically in a convergent competitive system, as each layer represents local arrangements of features in the preceding layer, and with each layer the representations become more complex with progressively larger receptive fields. Hence specificity of feature arrangements is built into the system from the very bottom, the system simply never has to address the binding problem.

However, not all the advantages of a sparsely coded output from the object recognition system can be appreciated without considering the representational significance of such output patterns. After all, in comparison with Churchland's account, which has complex objects being represented by vectors in fully distributed hyper-dimensional vector spaces, the present approach seems positively simplistic. Yet in this sort of system the representation is truly distributed, not just in a hidden layer, but amongst many layers. In this regard I am reminded of Dennett's Multiple Drafts model of consciousness:

> Feature detections or discriminations *only have to be made once*. That is, once a particular 'observation' of some feature has been made, by a specialized, localized portion of the brain, the information content thus fixed does not have to be sent somewhere else to be *re*discriminated by some 'master' discriminator.[46]

In a convergent architecture something is represented by a complex pattern of neuronal activation in one layer, but in the next layer it gets represented by a much simpler neuronal 'label', which then goes on to affect the further processing in the system in a way that respects the content of that complex pattern.[47] This applies to the relationship between all layers, but has a special significance for the output layer. The output is simple because of the sorts of systems it must interact with in order to produce behavioural advantages. The next section deals with the way convergent competitive networks interact with the other neural architectures described in this section to produce cognitive processing. This will allow a demonstration of the

---

[46] *Consciousness Explained*, p.113.

[47] One might hypothesize that the conscious sensory experience, the qualia, must involve the complex pattern in the earlier layers, in addition to the activity in the higher layers, so that consciousness would be envisaged to involve a large number of neuronal populations from many different regions. Which neuronal populations contribute to consciousness at any given moment probably depends on attention and the nature of the task being attempted.

advantages of real neuronal computation when compared with traditional connectionist models.

2.4 Interactions between Neuronal Populations and Brain Function

To appreciate the power of the various network architectures discussed in this chapter it is important to consider the computational problems that an organism faces, even once it has successfully recognized an object. After all, perceptual differentiation only becomes recognition when it grounds further advantageous cognitive processes. An organism must be able to judge the reward value of the object; should the object be approached or avoided? Additional benefit would be gained from being able to recall more specific details from previous encounters, if any, so that past experience could be brought to bear on present actions. As a concomitant of this there would also have to be a mechanism whereby the present occasion could be added to the body of data just mentioned. Of course it is obvious that what I am talking about here are emotion and memory. As mentioned in sections 2.1 and 2.2, these are grounded by pattern association networks and autoassociation networks respectively. It is the problem of interfacing with these types of networks that accounts for the nature and utility of the coding that takes place in the convergent competitive networks of the object recognition system, amongst other systems.

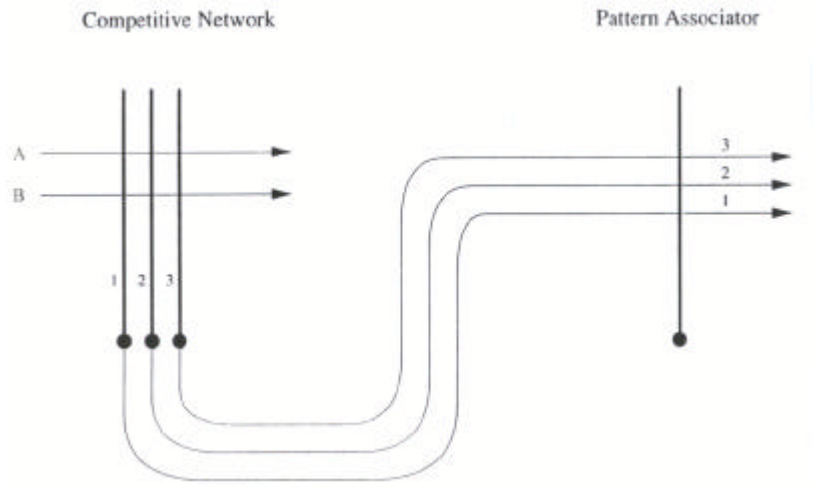

Figure 6: Expansion recoding. A competitive network is connected to a pattern association network to allow patterns that are not linearly separable (recall the notion of a hyperplane introduced in section 1.1) to be correctly learned.

2.4.1 Orthogonalization, Emotion, and Memory

Pattern association networks can be used as an alternative way of solving the exclusive OR problem, which was discussed in section 1.1. It is true that a pattern association network alone cannot solve the problem, because the Hebbian algorithm is not powerful enough. However, this can be overcome if the input to the pattern association network is pre-processed by a competitive network, as shown in Figure 6, on page 41. This is known as expansion recoding, because the original vector is transformed into a vector with increased dimensionality (i.e., it uses more neurons). Expansion recoding works because the competitive network orthogonalizes the patterns. Two vectors are completely orthogonal when they are at ninety degrees, at which point they are completely independent. Thus orthogonalization means that vectors are made less similar; the angle between them is made larger, approaching ninety degrees, thus reducing interference. With expansion recoding each input vector is represented by a separate output neuron, which allows the pattern association network to solve the problem with the weights shown in Table 3.

Recoded Inputs	Synaptic Weight
Input 1 (A=1, B=0)	1
Input 2 (A=0, B=1)	1
Input 3 (A=1, B=1)	0

Table 3: Weights required in the pattern association network following expansion recoding.

Thus where a standard connectionist model would solve this sort of problem with a hidden layer and a backpropagation algorithm, it seems plausible that the brain would use a combination of architectures. This sort of sophistication in categorization is needed in determining the emotional significance of objects, because similar objects may have very different reward values.[48] Thus objects receive different and orthogonalized representations which allows them to be more easily related with their reward value in pattern association networks. There is a large body of compelling evidence that these pattern associations take place in the amygdala and the orbitofrontal cortex. These areas receive inputs from the final stages of the object

[48] I am assuming that it is the evolutionary purpose of emotions to motivate appropriate behaviours towards objects and situations, i.e. on the basis of previous reward value.

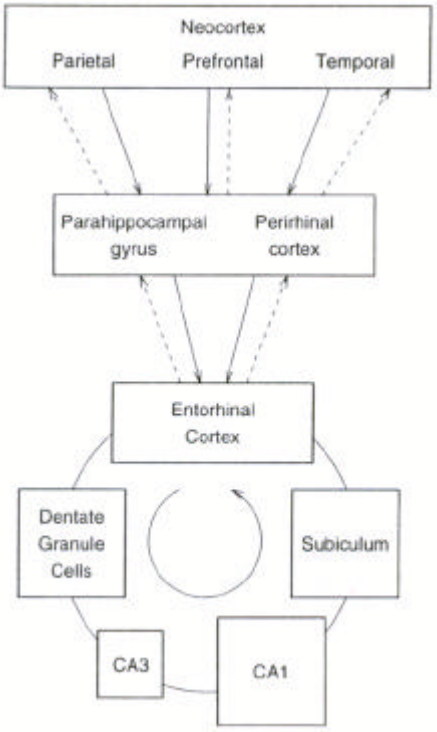

Figure 7: Schematic diagram of the connections in the hippocampal circuit. Forward connections are shown as solid, backprojections are shown as dashed lines.

recognition system in the inferior temporal cortex, but they also receive inputs from other sensory areas such as taste and olfaction. Their outputs include structures, such as the hypothalamus, that are involved in the control of autonomic functions which have a role in emotional responses. The amygdala is a subcortical structure involved in the more basic aspects of emotional behaviour in that single neurons do not show much flexibility in response to changes in reward value, where this can be thought of as the emotional significance of a stimulus. For example, food would have a positive reward value for a hungry animal. Neurons in the orbitofrontal cortex, on the other hand, do exhibit a close tracking of reward value even on single trials. The orbitofrontal cortex is part of the frontal cortex, the region of the brain that is most enlarged in humans when compared with other primates. The possible significance of this area for complex thought will emerge in section 2.5.

Orthogonalization plays a similar role for autoassociation networks. The role of episodic memory is to store information about particular occasions. Even if two situations are extremely similar they must be stored as separate temporal units. Marr was one of the first to suggest that the area responsible for this function might be the hippocampus, a subcortical structure in the temporal lobe, and the idea has been taken up by a number of others.[49] As shown in Figure 7, the hippocampus receives connections from many areas of the neocortex, via the entorhinal cortex. These are

[49] See D. Marr, 'Simple Memory: A Theory for Archicortex', *Philosophical Transactions of The Royal Society of London, Series B* 262 (1971), 23-81, J. L. McClelland, B. L. McNaughton, and R. C. O'Reilly, 'Why There are Complementary Learning Systems in the Hippocampus and Neocortex: Insights from the Successes and Failures of Connectionist Models of Learning and Memory', *Psychological Review* 102 (1995), 419-457, and E. T. Rolls, 'Parallel Distributed Processing in the Brain: Implications of the Functional Architecture of Neuronal Networks in the Hippocampus', in R. G. M. Morris, ed., *Parallel Distributed Processing: Implications for Psychology and Neurobiology* (Oxford University Press, Oxford, 1989), Ch. 12.

then passed through a circuit in the hippocampus before returning to the entorhinal cortex, and from there back to the neocortex. The crucial part of the circuit is the CA3 stage, which has a large number of recurrent connections, and thus seems a likely candidate for an autoassociation network. The idea is that the cortical activity in many areas of the brain during a particular experience pass patterns of activation into the hippocampus. The dentate granule cells act as a competitive network to remove redundancy and to orthogonalize the input so that large numbers of patterns can be stored without interference in much the same way as was explained for pattern association networks above. This pattern is then laid down in the CA3 network. Upon recall a fragment of the original pattern is presented to the CA3 network, where a complete pattern is produced. The next stage illustrates a very important principle of real neural functioning. The output pattern of the CA3 neurons is passed back to the entorhinal cortex, and from there to the neocortex so that the same, or at least a similar, pattern of activity is reproduced in the neocortex as was present when the memory was formed. The ramifications of this feature are discussed in section 2.5 below.

2.4.2 Sparsification

The sparseness of the representations produced by a competitive network is dependent on the degree of competition. The example of expansion recoding discussed in section 2.4.1 involved the greatest level of competition, in that it was a winner-takes-all network. It is unlikely that the competition employed by competitive networks in the brain would be of this kind, it is more likely to be soft competition as this presents computational advantages. The prime advantage is that sparse representations allow more patterns to be stored in pattern association networks and autoassociation networks. The mathematics behind this is complex, because the capacity of a network is dependent upon many factors, and is relative to the desired level of reliability of recall and how orthogonal the patterns are.

Sparse coding gives greater storage capacity in terms of number of patterns stored than either local or distributed representation, but still maintains the benefits of distributed encoding, such as generalization and graceful degradation. These benefits outweigh the cost of abandoning fully distributed encoding: sparsely coded patterns contain less information than their fully distributed counterparts, because less

representational elements are used. However, the amount of information still rises linearly with the number of neurons involved, and given that information is a logarithmic measure, this means that the representational capacity rises exponentially. Thus even a small number of neurons can encode a large number of patterns using sparse coding.

2.4.3 Recurrent Connections and Neural Processing

In explaining hippocampal functioning the recurrent projections to the neocortex played a crucial role, but recurrent projections are not unique to this circuit, they are ubiquitous throughout the brain. There are often as many, if not more, backprojections as there are in the forward direction. These backprojections, however, are not limited to the adjacent layer, they will often project to many other layers, and other networks altogether. Their very abundance suggests a key role in brain processing, and their role in the hippocampal memory circuit suggests one way in which they might function in other brain systems. The backprojections could act to reinstate original patterns of cortical activity during recall. Rolls and Treves give the following example:

> Consider the situation when in the visual system the sight of food is forward projected onto pyramidal cells in higher order cortex, and conjunctively there is a backprojected representation of the taste of the food from, for example, the amygdala or orbitofrontal cortex. Neurons which have conjunctive inputs from these two stimuli set up representations of both, so that later if only the taste representation is backprojected, then the visual neurons originally activated by the sight of that food will be activated. In this way many of the low-level details of the original visual stimulus might be recalled.[50]

Evidence for this comes from PET studies of subjects asked to recall visual scenes in the dark, which revealed increased blood flow, and thus increased neural activity, in early visual processing areas.[51]

Backprojections may also serve to aid learning in the cortex. Consider a competitive network that has a set of backprojections in addition to its normal input. These might come from the amygdala, representing emotional states. This modification of the network allows two similar inputs to be treated differently if they

---

[50] *Neural Networks and Brain Function* (Oxford University Press, Oxford, 1998), p. 243.

[51] S. M. Kosslyn, *Image and Brain: The Resolution of the Imagery Debate* (MIT Press, Cambridge, Mass., 1994).

are encountered in different emotional states, because the backprojections form an extra part of the input pattern. This mechanism allows the organism to learn fine discriminations between stimuli, guiding its category formation and the recognition of those categories.

In a similar way backprojections might form the basis for semantic priming, where activation of one semantic element increases the ability to respond to related elements. This might be done by backprojections producing a slight increase in the activations of those neurons which are involved in the recognition of the relevant semantic category. This would allow them to pass their thresholds more quickly, and thus to respond more rapidly (although not always correctly, there will always be a trade-off between speed and accuracy).

Backprojections could also be the basis for some mechanisms of attention and cognitive influences on perception, selectively facilitating the activity of relevant cortical populations. An instance of this would be situations in which one input pathway, say olfaction, has a noisy input. The olfactory pathway and the taste pathway both project to a cortical area where flavour is represented. The taste input could drive this higher cortical area into the correct pattern, and the backprojections from this area could then help the olfactory network to settle into the right pattern. In this way the conjunction of information from many channels can help to alter and 'clean up' the activation in the earlier neural populations.

2.5 General Principles of Brain Function

There are several general principles that can be drawn from the detailed nature of neural processing considered in this chapter. Signals between layers act like a lock and key system, for example: a competitive network can output a simple, orthogonalized, sparsely coded key, which can be easily stored by a pattern association or autoassociation network. During recall this key can then be reproduced to 'unlock' the original complex pattern of cortical activity. This suggests that complex tasks might be solved, even with simple learning algorithms, through the interaction of several different sorts of networks operating in concert. One might speculate that a large portion of the sophisticated behaviour which brains produce is the result of such interactions, possibly between many different networks. This fits

well with the fact, well known from the earliest study of the brain, that processing is carried out in multiple distinct regions, which are differentiated by their architectures and patterns of connectivity (recall the map of areas and connections in Figure 5 on page 37). This fact is one of the key sources of tension between connectionism and neurophysiology. It is understandable that connectionism, being a young field, with limited computing power available, should model simple single networks. Yet neurophysiology suggests that much of what is interesting in behaviour, both human and non-human, arises out of the interactions of many different cell populations. It may well be the case that these interactions are governed by entirely different principles than the local interactions between neurons in a single network. What are needed, then, are models of processing control mechanisms, or what Clark has called neural control hypotheses, in order to explain inter-network interactions.[52] Clark envisages these as structural features that are responsible for the control of other systems, gating the flow of information, and determining which type of activity is predominant at any time. One example of this sort of control mechanism is the reactivation of patterns of activity by backprojections, which was described in section 2.4.3, and so the lock and key model has a fundamental role to play here.

Damasio and Damasio have used backprojections as the basis for a model of knowledge representation.[53] At the heart of this model is the notion of a convergence zone, which is 'an ensemble of neurons within which many feedforward/feedback loops make contact'.[54] The purpose of the feedback connections is to activate patterns in many distinct cortical areas that are involved in earlier processing. These patterns would be reproductions of patterns that had occurred during previous experiences. The feedforward connections contact other convergence zones, so that there are hierarchies of convergence zones, and this is what makes the model new and interesting. Knowledge about basic features and individuals would be encoded by the activity in the lowest convergence zones, knowledge about entity categories at higher levels, with more complexity and generality being introduced with each step up the

---

[52] *Being There*, pp. 136-141.

[53] 'Cortical Systems for Retrieval of Concrete Knowledge: The Convergence Zone Framework', in C. Koch and J. Davis, ed., *Large-Scale Theories of the Brain* (MIT Press, Cambridge, Mass., 1994), pp. 61-74.

[54] 'Cortical Systems for Retrieval of Concrete Knowledge: The Convergence Zone Framework', p.71.

hierarchy. Thus simple features such as colours would require only a small amount of activity, whereas representations of more complex categories, such as animals, would require the activation of many more areas, including many subordinate convergence zones. Rolls' model of invariant object recognition can be seen as an example of such convergence zones. Indeed Damasio and Damasio hypothesize that the inferotemporal region is the locus for these convergence zones, based on the patterns of knowledge deficits that result from brain injury. Knowledge at the various levels is accessed through the activation of the appropriate convergence zone, which very rapidly reinstates the activity pattern in many disparate brain regions, and depending on the point in the hierarchy which is damaged, various categories of knowledge will no longer be available to the injured individual. Thus it is wrong to think of particular types of knowledge as being localized in particular brain regions, rather the access to that knowledge is controlled by a particular region.

Van Essen *et al.* have proposed a slightly different neural control hypothesis in which groups of control neurons gate the flow of activity from one population to another.[55] The control neurons act as a kind of attentional filter, given that the brain receives vast amounts of information from its sensory surfaces, and only has a limited processing capacity (Van Essen *et al.* estimate that only 0.1% of the information in the optic nerve can be processed at any one moment).[56] Whether the model they propose is accurate is a matter for further empirical research, but their analysis of the shortcomings of connectionist models is instructive:

> Conventional neural network models typically rely on computations that are dominated by linear combinations of synaptic feedforward inputs followed by a non-linear operation. This simple neural network structure has proven to be too rigid and unwieldy when applied to large problems . . . We suggest that models that do not distinguish control functions from information flow and processing will not scale well with increased problem complexity.[57]

Not only is there a need for attentional filtering, there is also a need to select between multiple processing pathways, given that there are many routes to output. For

---

[55] 'Dynamic Routing Strategies in Sensory, Motor, and Cognitive Processing' in C. Koch and J. Davis, ed., *Large-Scale Theories of the Brain* (MIT Press, Cambridge, Mass., 1994), pp. 271-299.

[56] 'Pattern Recognition, Attention, and Informational Bottlenecks in the Primate Visual System', *Proceedings of the SPIE Conference on Visual Information Processing: From Neurons to Chips* 1473 (1991), 17-28.

[57] 'Dynamic Routing Strategies in Sensory, Motor, and Cognitive Processing', p. 299.

example, an early processing layer may project to another processing layer, but it may also project directly to neurons involved in producing motor responses. This allows for both the production of fast, but stereotypical responses, and for slower but more complex and considered responses. The pathway that is most appropriate will depend upon situational factors, such as the general level of arousal, and so a mechanism is needed to make this decision. Such multiple routes to output may allow the same cognitive task to be achieved by several different processing strategies. One example of this is the existence of both a phoneme-based route to the production of speech from text, and a whole word recognition route. Evidence also comes from neurophysiology, where it has been shown that subcortical structures, such as the amygdala, have direct connections to motor neurons, as well as projections to higher neocortical structures.

The picture that emerges from analysis of real neural processing is one in which many distinct networks operate together. The principles whereby these interactions are controlled can only be guessed at, but they appear to be different in kind to the interactions that take place at a local level between neurons. An important point to make is that although control is needed, there is nothing in the hypotheses discussed above that suggests a central executive. The control mechanisms described have no access to the information stored in the systems they control. Some might argue that the correct style of explanation is of the traditional cognitive science 'black box' variety, but the fact that the models described are distinctly non-sequential in their connections speaks against this. However, a word of caution is needed. Many researchers agree that the frontal lobes play a crucial role in human cognition. Their major function seems to be planning, including response selection and suppression and mental modelling to predict real-world outcomes. Nothing in this section sheds any light on this directly. That such an important element of thought remains unexplained is frustrating, but only to be expected given its complexity. Whether mental planning can be explained by a neural control mechanism or by an approach involving more traditional processing modules and a central executive remains to be seen. I suspect that something like the former may be closer to reality. However, it may be the case that such dichotomies are misconceived, and that the truth is somewhere in between. What I have in mind here is a model in which there is some localization of processing role, but whose properties emerge from the dynamic nature

of the interaction between those modules, i.e., more a network of rivers and dams than black boxes.

This leaves the interesting question of where connectionist models should be placed in relation to these two styles of analysis. What the discussion in this chapter has revealed is that some localization seems inevitable. It is an established fact in neurophysiology that there are discrete regions with their own processing roles. The problem at issue is what one concludes from this. Despite all the claims to abstraction it seems as if connectionism bears the strongest resemblance to the neural goings on within the modules themselves, rather than to the mechanisms involved in controlling inter-network interactions. As a result I would predict that connectionism, in its current form, will prove inadequate at successfully modelling complex world-negotiating behaviour. This need not lead to a complete rejection of connectionism, however, because it may still explain a large proportion of or cognitive processing. In addition it may also adapt to take on the problem of larger scale interactions in such a way as to explain them in terms of emergent properties that does not fit with conventional artificial intelligence styles of explanation.

2.6 Summary

Philosophical expositions of connectionism have typically focused on feedforward network architectures, with hidden units, which are trained by the backpropagation learning algorithm. In this chapter, I have attempted to sketch out a rather different picture based on current knowledge of the architectures and functioning of real neural networks. The key points from this chapter that the reader will need to carry forward to the discussion in chapter 3 are:

- The distinctions in computational role between different architectures.
- The computational advantages to be gleaned from the interactions between different architectures.
- The importance of mechanisms to control the interactions of different networks.

3 Rethinking Vector Cognition

With these features of real neural processing in place I now want to suggest a number of ways in which Churchland's neurocomputational model might be adapted to avoid the difficulties raised in section 1.3. In that section I argued that contemporary connectionist models cannot be considered to display genuinely systematic, and thus symbolic, behaviour. A consequence of this is that connectionist networks can only be considered to have representations with non-conceptual content. Before tackling this problem I want to outline Terence Deacon's account of symbolic thought, as this introduces several notions that play a role in my proposed solutions to the problem of complex cognition and systematicity.[58]

3.1 Icon, Index, and Symbol

Deacon, borrowing from Peirce, sets out a scheme of three hierarchically arranged categories of referential association: *icon, index*, and *symbol*.[59] Iconic reference is the most basic form. An icon is usually thought to refer to its object through some form of physical similarity, but Deacon argues that this is not the foundation; the basis is rather 'that aspect of the interpretation process that does not differ from some other interpretive process.'[60] Thus it is taking something to be the same as something previously experienced, it is *recognition*. Physical similarity is the most obvious reason why one object is treated as iconic of another, but it need not be the only one on the definition given here. A picture of a person is iconic because of the stage in the interpretive recognition process which is the same for an actual encounter with that individual, or with the picture. In the present context iconicity can be conveniently assimilated to the vector prototype explanation of recognition; a stimulus is an icon of *x* if it activates the *x* prototype.

---

[58] *The Symbolic Species: The Co-Evolution of Language and the Human Brain* (Allen Lane, The Penguin Press, London, 1997).

[59] *The Symbolic Species*, p. 70; C. S. Peirce, 'Collected Papers. Volume II: Elements of Logic'. C. Hartshorne and P. Weiss, eds. (Belknap, Cambridge, Mass., 1978).

[60] *The Symbolic Species*, p. 76.

Indexical reference depends upon iconic reference, requiring the existence of at least three iconic relations. An indicating stimulus must be recognized as iconic of a previous class of stimuli. In addition members of this class must correlate with members of another class of stimuli which are seen as iconic of each other. Finally, and most importantly, these previous correlations must be interpreted as iconic of one another. This third relation is a higher-order icon, ranging over existing basic icons. An example of this type of reference would be the warning calls of vervet monkeys.[61] These monkeys produce distinct calls for different predators, such as snakes, eagles, and leopards. This involves the recognition of a predator, the selection and recognition of the correct warning cry, and the recognition of the previous correlations between the two. Given this explanation, indexical reference just seems to be another way of describing learned association of natural indicators.[62] Indeed a similar charge might be levelled at the account of iconicity, in that it is essentially perceptual recognition. Deacon raises just these questions:

> Could we just substitute the word 'perception' for 'icon' and 'learned association' for index? No. Icons and indices are not merely perception and learning, they refer to the *inferential* or *predictive* powers that are implicit in these neural processes. Representational relationships are not just these mechanisms, but a feature of their potential relationships to past, future, distant, or imaginary things. These things are not physically re-presented but only virtually re-presented by producing perceptual and learned responses like those that would be produced if they were present.[63]

Thus what is stressed in this account is the role of these processes in the cognitive economy. A sensory stimulus counts as an instance of an icon because of the way it is processed, as in the example of a picture, given above. It is not some feature of the stimulus which is the fundamental ground of reference, it is what is done with the stimulus. This matches well with the action-oriented nature of vector prototypes which was mentioned in section 1.2.

Symbols are at the top of the three-tiered hierarchy, as they range over indexical relationships. Deacon illustrates the nature of symbolic reference using the

---

[61] R. M. Seyfarth, D. L. Cheney, and P. Marler, 'Monkey Responses to Three Different Alarm Calls: Evidence for Predator Classification and Semantic Communication', *Science* 210 (1980), 801-3.

[62] For an exposition of the notion of 'indicator aboutness' see F. Dretske, 'Misrepresentation', in R. Bogdan, ed., *Belief: Form, Content, and Function* (Clarendon Press, Oxford, 1986), pp. 17-36.

[63] *The Symbolic Species*, p. 78.

example of a set of chimpanzees who were trained to use symbolic communication.[64] The chimps were taught to use a computer keyboard with simple abstract shapes known as lexigrams on the keys. Previous experiments had shown that chimps were capable of learning a large number of lexigram-object associations, that is, indexical relations. However, this does not constitute symbol-use, and the reason why it does not provides a clue to the features which are essential to symbolic relations. If the lexigram and the object are no longer paired and rewarded, the association will be extinguished; but words are only rarely paired with their referent, and so indexical reference is problematic as the sole basis for full-blooded reference. This is the central weakness in causal theories of reference.

Deacon argues that the missing element in indexical reference is syntax.[65] An attempt was made to train the chimps to comprehend a simple syntactical system which involved a simple verb-noun relationship. The two 'verbs', one for solid food, and one for liquid, had to be paired with an appropriate noun, 'banana' for instance, in order to get the item. The chimps only managed to master this simple system through a long and highly structured training regime, which cued them to both relevant and irrelevant features. Mastery of the system was tested by comparing the speed with which they grasped new lexigram 'nouns' against non-symbolically trained chimps. It was found that the specially trained chimps learned the function of the new lexigrams on their first presentation, or after only a few trials, whereas the control chimps took hundreds of trials, as usual. This illustrates both the nature and advantage of symbolic representation:

> What the animals had learned was not only a set of specific associations between lexigrams and objects or events. They had also learned a set of logical relationships *between the lexigrams*, relationships of exclusions and inclusion. More importantly, these lexigram-lexigram relationships formed a complete system in which each allowable or forbidden co-occurrence of lexigrams in the same string (and therefore each allowable or forbidden of one lexigram for another) was defined. They had discovered that the relationship that a lexigram has to an object *is a function of* the relationship it has to other

[64] See D. Rumbaugh, ed., *Language Learning by a Chimpanzee: The Lana Project* (Academic Press, New York, 1977).

[65] An argument to support this reliance on syntax and combinatorial possibilities as the basis for genuine symbolic meaning will be given in section 3.3.

lexigrams, not just a function of the correlated appearance of both lexigram and object. This is the essence of a symbolic relationship.[66]

Thus symbols are higher order categories of indexical relationships. Lexigrams are recognized by which category they belong to, either verb, which might be thought of as 'give', or noun. Members of each of these categories has a fully determined set of combinatorial possibilities with other tokens, dependent upon which category they fall into in turn. Hence new lexigrams can be acquired rapidly, because once it is established which category they belong to their role in the system is grasped, there is no need to learn their associations from the scratch. These logical categories are defined by their combinatorial possibilities, and so an entire system of interrelations must exist before any single token can be considered as a symbol.

A very special training regime was necessary in order to get the chimps to acquire this rudimentary symbol system successfully. The transition from indexical reference to symbolic reference requires the recognition of global patterns amongst a large number of lexigram-object associations. This requires a change in perspective of, what was for the chimpanzees, a previously acquired body of indexical knowledge. Progress is further hampered by the fact that symbols from the same category will not appear together, indeed they will occur in the context of symbols from different categories. However, once this step has been taken it allows a considerable off-loading of cognitive effort. One only has a limited number of interdependent categories to recognize, rather than a huge array of independent associations. Symbolic reference is also more powerful because it moves beyond simple naming functions. Indices are grouped together because of a similarity of relationships between indexical token and object, and this linking of symbols to relationships changes the focus from objects to classes of relationships between objects, allowing for more complex representations. Thus once a basic symbol system has evolved it is possible for more complex operations to be added:

The system of representational relationships, which develops between symbols as symbol systems grow, comprises an ever more complex matrix. In abstract terms, this is a kind of tangled hierarchic network of nodes and connections that defines a vast and constantly changing semantic space.[67]

---

[66] *The Symbolic Species*, p. 86.

There is nothing new in the idea that combinatorial syntax is crucial to symbol-use and language. It played a central role in the rise of modern logic with the work of Frege and Russell, and the early Wittgenstein. It is also involved in many modern philosophical accounts of concepts, including Fodor's, the patriarch of symbolic models of cognition. Further, and somewhat paradoxically, Deacon's account bears some relation to conceptual role semantics, in that symbolic status depends upon relations to other symbols. Thus the syntactic and semantic categories of an item are determined by the pattern of its connections with other symbols. However, these connections are not the sole determinants of content, the process of comprehending symbols is viewed as moving in a downward direction in the representational hierarchy, from symbol, to index, to icon. Production of symbols involves a move in the opposite direction. This compositional relationship gives semantic content to symbols. They are not merely syntactic entities, they are composed of perceptual and action-based processes as well; and these are the very same processes that are deployed in the cognition of non-symbolic species.

3.2 Neural Networks and Symbols

It is important to be clear about the exact way in which Deacon's account is relevant to the present discussion. The example of symbol acquisition in chimpanzees may not be totally analogous to the situation of the human infant. It is open to speculation whether neonates first learn iconic relationships, then lexical relationships, before having a eureka experience in which these elements are reorganized. I am construing him as making a claim about the way in which we should analyse symbolic thought, rather than as making a claim about the ontological dependence in development between referential relations. The important aspect of Deacon's analysis is the way it is constructed from elements of which we have a (relatively) comprehensive neurophysiological understanding: recognition and associative learning, in the guise of iconicity and indexicality. Iconic reference can be understood as being based upon object prototypes, implemented in the form of convergence zones. The mapping of an

---

[67] *The Symbolic Species*, p. 100. This could be linked in an interesting way with the work of Adrian Cussins, particularly his idea of cognitive trails and the rise of objectivity and perspective independence, which could also be related to the discussion in this section, see A. Cussins, 'Content, Embodiment and Objectivity: The Theory of Cognitive Trails', *Mind* 101, 651-88.

input onto a convergence zone in this system would constitute recognition. Whilst words (whether spoken or written) are indexical tokens, thus having relations to icons, they are also symbols, and so they have something of a double life. I now want to suggest how these final levels in the hierarchy — from words to symbols — might be instantiated by a parallel processing system, and how this might go some way to solving the problem of systematicity that was raised in section 1.3.

Psycholinguistic models suggest how lexical tokens might produce symbolic thought. An important feature of human language comprehension is the way that spoken words are recognized.[68] As phonemes are received the words which are compatible with them are activated, but as more phonemes are heard it appears that more and more words drop out of contention as they become inconsistent with the auditory input. Eventually there is only one left, and this is the word that is recognized. For example, for the word *kangaroo* the word is identified as soon as the phoneme /g/ has been heard, since at that point no other word is consistent with this input. This cohort model of lexical access has been reproduced by D. Norris in a recurrent network based on the one designed by Elman, that was discussed in section 1.1.[69] Remember that Elman's network was tested by being asked to predict the following word in a sentence. Its response was to activate all words that were compatible with the initial part of the sentence. Norris' network had 50 output units, each one representing a word. The network was trained by being presented with a phoneme in each processing cycle, with no breaks between words, just as in real speech. In testing, single words were presented and the activation levels of the words were examined. The cohort pattern was found with words dropping in activation when inconsistent phonemes were presented, until only the winning word unit was left. Interestingly the network exhibited many features of human word comprehension, such as patterns of identification of mispronounced words. For instance, humans can correct for some mispronunciations of the first phoneme of a word if they are close enough to the intended target. The network could recognize

---

[68] Although there are significant differences in the case of written word comprehension, the overall strategy appears to be the same for, and the differences are not important in the present circumstances.

[69] D. Norris, 'A Dynamic-Net Model of Human Speech Recognition', in G. T. M. Altmann, ed., *Cognitive Models of Speech Processing: Psycholinguistic and Computational Perspectives* (MIT Press, Cambridge, Mass., 1990), pp. 87-104.

goronet as a distortion of *coronet*, but not *horonet*, as the /g/ is close enough to /c/, but the /h/ is not.

Thus recurrent network architectures can explain how words are recognised, but they can also cope with grammar. Elman trained another network on a complex grammatical task, that was much more challenging than lexical-category structure. The task involved a lexicon of 23 items, including 12 verbs and 8 nouns, and a phase-structure grammar. This meant that the network had to be able to cope with recursive structure and complex relationships, a truly rigorous test of representational capacity. The inputs were again local, each unit representing a word, and one might envisage a situation in which a word recognition network could provide such local outputs, which could then act as inputs to the grammatical network. The training of the grammatical network was done in phases, starting with simple sentences and gradually increasing the number of complex sentences over time.[70] The first analysis of the network was similar to that for the original lexical-category structure network: words were presented and the network had to predict the next word. As with the more basic network there is no way to tell exactly which word would come next, rather in this case the following word had to be from a grammatically correct category. Surprisingly, the network performed very well at this task, grasping even the most subtle and complex relations. Elman gives the sentence 'Boys who Mary chases feed cats' as an example of this ability:

> The appearance of *boys* followed by a relative clause containing a different subject (who Mary) primes the network to expect that the verb which follows must be of the class that requires a direct object precisely because a direct-object filler has already appeared. In other words, the network not only correctly responds to the presence of a filler (*boys*) by knowing where to expect a gap (following *chases*); it also learns that when this filler corresponds to the object position in the relative clause, a verb that has the appropriate argument structure is required.[71]

---

[70] This feature of the experimental design, in which the network was started with simple inputs which then became progressively more complex, has been stressed in the context of developmental plausibility. It is argued that the short attention spans and biases of infants might act in the same way, and that this might significantly simplify the task of comprehending the complex relations that grammar allows, see J. L. Elman, E. A. Bates, M. H. Johnson, A. Karmiloff-Smith, D. Parisi, and K. Plunkett, *Rethinking Innateness: A Connectionist Perspective on Development* (MIT Press, Cambridge, Mass., 1997), pp. 340-349, and also T. Deacon, *The Symbolic Species*, Ch. 4.

[71] 'Representation and Structure in Connectionist Models', p. 363.

The network could cope with several nested centre embeddings, although its performance decreased with each one, in a manner similar to humans. This suggests that the network has a genuinely productive ability that is only limited by computational resources. The sophistication of these grammatical abilities is impressive, but it leaves the question of how the network has managed the task. The advantage of a recurrent network is that it can take temporally extended inputs and produce temporally extended outputs, but the disadvantage is that this makes analysis extremely difficult, as Elman explains:

> In the previous simulation [the lexical-category structure network], hierarchical clustering was used to reveal the use of spatial organization at the hidden-unit level for categorization purposes. However, the clustering technique makes it difficult to see patterns that exist over time. Some states may have significance not simply in terms of their similarity to other states but also with regard to how they constrain movement into subsequent state space . . . Because clustering ignores the temporal information, it hides this information. It is more useful to follow over time the trajectories through state space that correspond to the internal representations evoked at the hidden-unit layer as a network processes a given sentence.[72]

Unfortunately it is virtually impossible to visualize a trajectory in a hyper-dimensional vector space, so a method called principal component analysis has to be used to reveal those hyperplanes in the hidden unit vector space that involve most variance. This reveals that similar grammatical structures are represented by similar trajectories in vector space; but as with spatial semantic metric revealed by cluster analysis, slight differences were marked by slight divergences in trajectory. The most crucial point to emerge from this is that movement through recurrent vector space is constrained; at any point there are only a certain number of trajectories available, and these encode the grammatical progressions. Just as simple feedforward networks were explained in terms of attractors in vector space in section 1.1, so recurrent networks can be understood in terms of well-worn paths in vector space. This explains how the network is able to activate only words that fall into appropriate grammatical categories during the testing phase. Elman's network is not an accurate model of how the brain processes grammar, the use of backpropagation sees to this, but biological plausibility is not the purpose of this model. Rather the importance of Elman's network is that it reveals how certain structural features — recurrent connections —

---

[72] 'Representation and Structure in Connectionist Models', pp. 363-4.

make it possible for connectionist networks to represent the combinatorial possibilities of symbols. This is a capacity that has been denied by opponents of connectionism. The hope must be that the principles are right and that what remains is to discover how they can be implemented in real neural systems.

3.3 Symbols and Semantic Content

So far no explicit explanation has been given of semantic content. Elman's networks only processed grammatical and lexical relationships, and so without further explanation this is just a meaningless kaleidoscope of activity, in the same way that the voltage fluctuations in the microchips of a computer are without an external interpretation. After all, a point in vector space might represent a word in a particular grammatical role, but how does this come to mean anything? There are three elements in Deacon's account that contribute to the explanation of meaning, as demonstrated in the following passage:

> [The] cross-modal associations [in convergence zones] between images and experiences on the one hand and their associations with particular word sounds on the other provide the indexical associations of words, but their symbolic association — what we call the meaning — involves these and something more. The something more includes both the associative relationships between words and the logic of how these map to the more concrete indexical relationships.

Thus the first element is the dependence of symbolic status upon combinatorial possibilities with other symbols, the second is the associations between words, and the third is the indexical relations of symbol tokens. I will combine this view with the foregoing discussion of neural computation in sections 3.3.1, and 3.3.2.

3.3.1 Context and Meaning

An account of meaning can be extrapolated from Churchland's vector prototype model of explanatory understanding. Recall that the basis of this model is that a prototype represents an object because it comes to encode the contexts in which that object has been experienced previously. Translating this to language, the pattern of neural activity that represents the meaning of a word, a concept, does so because it reflects the contexts in which that word has occurred. The notion of context here is a rich one, including not just linguistic context, but also other elements of experiential context. Elman's grammatical network could predict which classes of words would

come next in a sentence, thus at a very rudimentary level it had encoded the contexts of words. Of course, one consequence of this is that there is no strict separation of syntactic and semantic aspects of processing. All aspects of context that are predictive will be incorporated in the configuration of weights. This is an advantage because semantic cues can be used in lexical processing. That humans do this sort of thing can be demonstrated by experiment. Altmann discuses the following example:[73]

Which *woman* did Bertie present a wedding ring to?
Which *horse* did Bertie present a wedding ring to?

As soon as the word 'present' is heard it can be given two alternative interpretations, the woman or horse is either the thing being presented, or the recipient of the object being presented, whatever that turns out to be. When 'wedding ring' is heard this suggests that it is the thing being presented, and so the horse or woman must be the recipient. A horse is an implausible thing to give a wedding ring to, but if we waited until the end of the passage before doing the grammatical processing, we would not notice this implausibility until that point. However, EEG recordings show that subjects notice the implausibility when 'wedding ring' is heard, not when the end of the passage is reached. Although this is a very short time, it is nonetheless significant, and shows that roles are assigned as soon as they are registered, and that semantic factors play a role in that determination. We assume that wedding rings are the sorts of things that get presented to people, and use this piece of knowledge in the analysis of grammatical structure even when what could follow might disagree with these assignments. For instance, the passage above might continue as follows:

Which woman did Bertie present a wedding ring to his fiancée in front of ?

This sentence is difficult to process precisely because it conflicts with the assignments we are inclined to give; we assume that the wedding ring is being given to the woman. Such semantic cues will not be as influential as more syntactic features, given that on the whole they will be less predictive. Fortunately counterintuitive sentences like the one given above are the exception rather than the norm.

---

[73] *The Ascent of Babel: An Exploration of Language, Mind, and Understanding* (Oxford University Press, Oxford, 1997), p. 110; the work discussed was carried out by M. K. Tanenhaus, J. E. Boland, G. N. Mauner, and G. Carlson, 'More on Combinatory Lexical Information: Thematic Effects in Parsing and Interpretation', in G. T. M. Altmann and R. C. Shylock, eds., *Cognitive Models of Speech Processing: The Second Sperlonga Meeting* (Lawrence Erlbaum Associates, Hove, 1993), pp. 297-319.

This analysis of meaning owes something to Wittgenstein, for several reasons. First, Wittgenstein propounded the idea that an explanation of meaning should proceed via an investigation of use, and as demonstrated above, the significance of a symbol springs from its relations with other symbols, and its linguistic contexts, which are reflected in its use. Secondly, it was Wittgenstein's insight that meaning does not have to be exact ('everywhere bounded by rules'[74]) in order to function satisfactorily. Given the nature of symbols, and the evolutionary forces which have shaped language, it comes as no surprise that a precise axiomatization is not a prerequisite for language, as will be argued in the next few paragraphs, and in section 3.4. I will attempt to produce an analysis of meaning that is linked to what has been learned about its neural substrates. This kind of approach would have been anathema to Wittgenstein, as is demonstrated by the following passage:

> No supposition seems to me more natural than that there is no process in the brain correlated with associating or thinking; so that it would be impossible to read off thought-processes from brain-processes. I mean this: if I talk or write there is, I assume, a system of impulses going out from my brain and correlated with my spoken or written thoughts. But why should the *system* continue further in the direction of the centre? Why should this order not proceed, so to speak, out of chaos? The case would be like the following — certain kinds of plants multiply by seed, so that a seed always produces a plant of the same kind as that from which it was produced — but *nothing* in the seed corresponds to the plant which comes from it; so that it is impossible to infer the properties or structure of the plant from those of the seed that comes out of it — this can only be done from the *history* of the seed. So an organism might come into being even out of something quite amorphous, as it were causelessly; and there is no reason why this should not really hold for our thoughts, and hence for our talking and writing.[75]

I think that Wittgenstein's position is unsatisfactory in its rejection of any investigation of causal underpinnings for language. His analogy with a seed reveals how bizarre his position is; I cannot help but think of the development of a seed in terms of cellular mechanisms and genetics. To say that these do not in any way bear on its eventual structure is just untenable, and unscientific — it is the opposite of scientism. The neurally inspired model of cognition offers a richer account of linguistic behaviour and its underlying cognitive mechanisms that I believe occupies a midway position between these defective extremes.

---

[74] *Philosophical Investigations*, §84.

[75] *Zettel*, § 608.

The vector realization of meaning bears a family resemblance to functional role theories, but the incorporation of context by degree of predictiveness allows a response to the standard criticism that such theories lead to no sharp distinction between empirical and semantic aspects of concepts. This charge can be accepted without making the theory unattractive and unrealizable. I am inclined to deny that there are any inferential roles which have a privileged status in determining meaning. There will probably be some which are more important than others, and this will be reflected in the fact that they will be shared by most language-users. This in turn will be due to the fact they are the most predictive aspects of context. Thus the more important inferential roles will be determined by the nature of the underlying processing system, with its ability to respond appropriately to statistical tendencies in the environment. Hence the argument that functional role theories make sameness of meaning impossible can also be parried. As long as two people have neural patterns that are activated in more or less the same situations, then those patterns can be considered to have the same meanings for them. Obviously this makes meaning a matter of degree, because we could not expect total co-activation in every possible circumstance, but this need not be seen as detrimental. One might argue that it allows for the idiosyncrasies of individual experience to be reflected in cognitive life. A traditional difficulty with functional role theories is how they connect to the world of objects. The solution to this problem lies in the hierarchical component of meaning, that Deacon rightly emphasizes, and this is where indexical and iconic reference re-enter the picture.

3.3.2 Meaning and Indexical Foundations

Connections between symbol tokens mean nothing if they are not properly grounded in experience. This grounding is achieved through the interaction of many different networks in the brain, especially convergence zones, that underlie the functioning of iconic and indexical reference. At the simplest level there are links between words and objects. Such word prototypes, activated in a grammatical network, might cause activation of the appropriate convergence zone. This provides the content to the symbol, the convergence zone goes on to rapidly activate patterns in a variety of brain areas. These reinstated patterns have content because of the brute fact that they played the same role during the original experience. For example, at the most basic level, perceptions of colour are localized in a particular region of the visual cortex. How

this can be is, perhaps, inexplicable as far as human investigation is concerned. I have made some comments about the remarkable representational power of vector coding, but for present purposes it must be accepted as brute fact that these collections of neurons produce an experience of colour. This explains the content of basic perceptual features. The next stage is the conjunction of these into representations of objects at the next level in the convergence zone hierarchy, and then into categories of objects, and then superordinate categories, and so on. It is likely that the train of processing will not stop at this point, as the pattern of activity caused by the activation of the convergence zone can generate further activity. For example, one might hear someone say the name 'John', which would activate one's convergence zone for that individual, and this in turn might stimulate the recall of information about them, and of past encounters with them through the operation of autoassociative networks. The exact nature of this cascade of activity would depend on the cognitive context in which the name is heard. First, if one knew several people with this name, then context would serve, in most cases, to activate the correct convergence zone. Second, which episodes would be recalled would depend upon the context in which they were mentioned. Such contextual factors might be things as general as mood, as it has been found that it is easier to recall happy thoughts when in a positive mood, and visa versa for depressive thoughts. Alternatively, contextual factors might be highly specific: considering the suitability of candidates for a given task would trigger recall of past performances. It is this rich pageant of ongoing activity that accounts for the complexity and richness of human cognitive experience.

3.4 Symbols, Systematicity, and Concepts

The foregoing account of symbolic processing and meaning goes some way to solving the problem of systematicity. However, recall that in section 1.3, I argued that having a complex semantic metric in the hidden layer of a recurrent network is not enough for true systematicity. The reason for this was the inability of such networks to use their knowledge in performing different tasks, their content is non-conceptual. As they stand, then, Elman's networks do not exhibit systematicity. But if the output of these networks was sparsely coded, it could be input to another network, which could then perform other operations upon it. Thus activation in the verb part of the hidden layer vector space could cause a certain sparse label to be produced, and this

could allow a lexical category decision to be made by a separate network. The very same information could be made available to a number of networks in this way, each one being used in a different cognitive context. The selection between these different processing routes would probably need to be controlled in the sort of way mentioned in section 2.5, although in some situations there might be competition and mutual inhibition of the sort that occurs in competitive networks (see section 2.3). This analysis of systematicity would have the consequence that activity in the original network would be non-conceptual, whilst it would be made fully conceptual through interaction with other networks. Hence the notion of 'concept' is one that can only really be applied at the level of complete systems (this reflects the intuition that it is a personal, rather than sub-personal, notion). Further this makes it possible to explain how one might have degrees of objectivity and perspective independence: the more complex and numerous the systems that the information is available to the nearer cognitive behaviour approaches to the ideal encapsulated in the generality constraint.

Sparse coding also opens up the possibility of several different patterns in a network being activate at the same time, suggesting that localist microfeatures could be utilized. In a fully distributed system, the activity of every element is involved in the representation of any single item. If the coding is much more sparse it might be possible for two objects to be represented simultaneously if their patterns did not involve the same units, i.e., if they were sufficiently dissimilar. This might make it possible for several elements that are stored it the same network to be represented at the same time, thus overcoming another objection to the systematicity of connectionist models, and recurrence provides the capacity to represent relations between these sparse representations.

Churchland's model of conceptual redeployment, which was discussed in section 1.2.2, also offers a possible explanation of systematicity. For it goes some way to explaining how a piece of knowledge gained in one situation can be applied in another. At that point I raised some doubts about how Churchland's idea might work, but the convergence zone hypothesis and the hierarchical nature of processing suggest how it might be achieved in the brain. A change in which a higher order convergence zone is activated can cause activity in more basic levels to be arranged in a different way, as different convergence zones activate different patterns of activity in different cortical areas. Thus in the case of the duck/rabbit, basic visual features such as edges

and lines might be mapped onto different features at higher levels, such as the area where form is processed. So the very same visual input can be interpreted as either iconic of a rabbit's ear or a duck's bill. This reinterpretation would require alterations throughout the system, but this can easily be achieved within the convergence zone framework. This framework might explain why it is extremely difficult, if not impossible, to visualize two different scenes simultaneously, the two patterns cannot be activated on the same units at the same time.

When this idea is applied to more cognitive phenomena it suggests another possible source for the systematicity of thought. The activation of different higher level prototypes might constitute conceptual redeployment by reorganizing the interpretation of sensory information. The convergence zone framework was devised to explain concrete knowledge, not the sort of knowledge involved in theoretical understanding. What I am trying to do here is extrapolate from those systems that we are beginning to understand. The spatial and temporal semantic metric of a multidimensional vector space explains how a point in that space can have representational content. However, such a space alone suggests no obvious mechanism for how activating one prototype can affect so many various aspects of cognition. But in hierarchically organised systems it might be possible for some hybrid model to accrue the benefits of both systems.

3.4.1 Neural Commitments

This account of systematicity does not yet provide a total explanation of the special nature of concepts. There is one more important point that must be added before the explanation is complete. I have admitted that the generality constraint does state a truth about concepts, but the important question is what follows from it. One source of conflict concerns its consequences for the cognitive underpinnings of concepts. Do all instances of a concept require some sort of causal commonality, and can a connectionist account provide the goods? It is clear that Evans did not mean for the generality constraint to lead straight into the adoption of a language of thought:

> . . . I certainly do not wish to be committed to the idea that having thoughts involves the subject's using, manipulating, or apprehending *symbols*—which would be entities with non-semantic as well as semantic properties, so that the idea I am trying to explain would amount to the idea that different episodes of thinking can involve the same symbols, identified by their semantic and non-semantic properties. I should prefer to explain the sense in which thoughts are structured, not in terms of their

being composed of several distinct elements, but in terms of their being a complex of the exercise of several distinct conceptual *abilities*.[76]

To say that different occasions are to be unified by the operation of the same ability necessitates an account of how abilities are individuated, and their place in the ontological scheme of things. Evans argues that there should be a 'common explanation' for applications of the same ability:

> Each common explanation will centre upon a state — the subject's understanding of '*a*', or his understanding '*F*' [for an explanation of the thought that *a* is *F*] — which originated in a definite way, and which is capable of disappearing (an occurrence which would selectively affect his ability to understand all sentences containing '*a*', or all sentences containing '*F*').[77]

I think it is right that there should be *some* commonalties across occasions, but not a complete similarity in all cases — no state present in every instance (in contrast to traditional AI models). What I have in mind here is that in different contexts we should expect different neural substrates to be used, given the multiple network nature of neural processing that was sketched out above. Relevant contextual aspects might include the kind of sensory input, e.g., a word, or a visual presentation of an object, and the kind of action being attempted, from visual search to solving a crossword puzzle, and so on. Evidence for this style of processing comes from PET and MRI studies, which have shown that different tasks recruit different brain regions. For example, in one study subjects had to listen to a story and monitor for either grammatical errors or for words in a particular semantic category. These activities caused activation in separate, but overlapping, areas of the ventral prefrontal cortex.[78] This type of evidence from brain-imaging studies indicates that there is probably significant task decomposition in the brain. It may be that certain networks are given over to the detection and processing of certain grammatical features. Examples might include the generation of words from the same semantic category, or alternatively words from complementary grammatical categories, such as an appropriate verb for a noun. One might envisage hierarchies of processing stages in which first the overall structural elements are identified, so that they can be processed

---

[76] *Varieties of Reference*, pp. 100-1.

[77] *Varieties of Reference*, pp. 101-2.

[78] P. J. G. Nichelli, J. Grafman, P. Pietrini, K. Clark, K. Y. Lee, and R. Miletich, 'Where the Brain Appreciates the Moral of a Story', *NeuroReport* 6 (1995), 2309-2313.

by more specialized networks. Evidence for this hierarchic approach comes from studies of cortical stimulation on conscious patients undergoing brain surgery.[79] Deacon has summarized the findings as follows:

> What these stimulation studies demonstrate is that the regions where stimulation disrupts language function fan out from the frontal mouth area into the prefrontal lobes, and from around the auditory area back into the temporal and parietal areas. Those regions where stimulation reliably disrupts the same language functions are organized in what appear to be tiers radiating outward from these two foci. Electrical stimulation of the regions closest to the motor and auditory areas produces problems with phoneme identification and oral movements. Stimulation further out disrupts naming of familiar objects and grammatical assessments. And stimulation even further out appears to disrupt retention or recall of words. There is also a rough front-back mirror symmetry of these tiers, so that the very same responses are elicited by the second and third tiers both front and back.[80]

It follows that in similar contexts the same networks would be used and so there would be limited commonality. What gathers these varying contexts together is the personal level attribution, made solely on the grounds of behaviour, linguistic or otherwise, rather than a computational commonality. Of course in most cases of philosophical interest the context will be one of language comprehension, and so a certain degree of uniformity is guaranteed, given the account of meaning given in section 3.3. If the same concept is being exercised in comprehending sentences, then the same neural pattern must be active, because this is what encodes predictive context, and therefore meaning. However, we should not let the apparent ease with which we can group different situations under the head of a given concept lull us into thinking that every exercise of a concept must have an immutable core of representative neural encoding. Rather some elements of neural activity will prove more central than others.

Is this a satisfactory way to philosophically unpack the notion of a 'common explanation'? It seems right that there should be a demand for some kind of causal commonality behind conceptual attributions — if we found that there was not intuition suggests that we might be persuaded to withdraw our attributions. The

---

[79] See, Penfield and Roberts *Speech and Brain Mechanisms* (Oxford University Press, London, 1959), G. A. Ojemann, 'Cortical Organization of Language', *Journal of Neuroscience* 11 (1991), 2281-2287, and G. A. Ojemann and C. C. Mateer 'Human Language Cortex: Localization of Memory, Syntax, and Sequential Motor-Phoneme Identification Systems', *Science* 205 (1979), 1401-1403.

[80] *Symbolic Species*, p. 289.

criterion, then, is whether one can provide strong enough causal explanations to defeat such challenges. Even though, on the multiple network model, it isn't possible to move from details of neural processing to personal level determinations of conceptual content in any principled fashion, there is nevertheless a relationship between these levels. I would argue that the similarities in neural activity across occasions that were postulated above would prove enough to meet the criterion.

Further support for this analysis comes from considering the exact nature of the personal level account. Is it clear that on all occasions there has to be a precise matter of fact about which concepts are exercised? An analogy from Geach demonstrates how it might be wrong to think that this is the case:

> The exercise of a given concept in an act of judgement is not in general a definite, uniform sort of mental act; it does not even make sense to ask just how many concepts are exercised in a given judgement. Our chess analogy may here again be of service, in showing why this question is unreasonable. Playing chess involves a number of abilities, which are not only distinguishable but can actually exist separately; for one way of teaching chess would be to play first just with the kings and the pawns and then add the other pieces successively in later games. It would, however, be absurd to ask just how many of these abilities there were, or just how many were exercised in a particular move; although one might perfectly well say that somebody knew the knight's move, and that this knowledge was or was not exercised in a particular move.[81]

This suggests the right way to analyse abilities: normally one has a competence in a task domain; one can focus on elements of that competence, and call them specific abilities. In this way abilities are identified from an external perspective, one picks out a certain subtask and investigates whether an individual can accomplish it, if so they can be ascribed the ability. However, it makes no sense to think that abilities operate independently in normal negotiation of the task domain. Rather, the whole complex system faces the task domain. The neurally inspired account of symbolic thought suggests a possible way to apply this chess analogy to language: a subject's symbolic abilities can be considered to be grounded in a number of networks, including recurrent grammatical networks, with their multidimensional vector spaces, and in neuronal populations that contain convergence zones. The basis for this claim would be that these neuronal populations ground the many abilities that a genuine symbol-user must exhibit. Full-blooded language comprehension might be a matter of

---

[81] *Mental Acts*, p.15.

aural or visual input being mapped through word recognition networks into grammatical networks, which then spread activation to the appropriate convergence zone prototypes, and other systems, providing the semantic content to the symbols, so that they are not empty syntactic shells as in a language of thought architecture (although advocates of the Language of Thought would obviously argue the point here).

However, each symbol cannot exist as a discrete entity, and so the chess analogy is not quite right. As made clear in section 3.1, a symbol depends for its identity upon its relationships to other symbols, so single symbols cannot exist, rather a simple language core forms the initial base, which can then be embellished with the addition of further grammatical categories. This core is defined by the combinatorial possibilities between its elements, and so in some sense a single symbol cannot be exercised without the existence of others. In practice this holism is realized in the grammatical vector space of Elman's recurrent networks, because they are trained on whole sentences, and only in this way do they learn interrelationships between words. A particular point in vector space constitutes a given symbol because of the paths which move off from it. If those other moves could not be made then the individual could not be ascribed that symbolic content. Thus a passage of cognition might involve a whole body of expertise, even though it is possible to be definite about which symbols were involved. The neural system grounds a symbolic system, not individual symbols.

So far I have used the terms 'symbol' and 'concept' interchangeably, without comment. As far as I am concerned they are virtually synonymous notions, at least that is how I treat them; the only difference is in philosophical connotation. 'Concept' is traditionally linked with sensitivity to public agreement and rule-following. In Wittgensteinian parlance, logically private concepts, that cannot possibly be shared, are a philosophical nonsense. In recognition of this I will define a concept as a rule-governed ability. Adopting this notion allows the neural account of cognition to both satisfactorily explain the physiological basis of a concept-user's abilities whilst simultaneously avoiding a reduction of concept-use to any kind of causal regularities. Thus allowing us to avoid the dichotomy that McDowell sets up between empirical and rational styles of explanation; the normativity of concepts is properly

acknowledged.[82] As argued above, a concept is constituted by many varied abilities, and competence can be ascertained from an external perspective, thus delivery the required objectivity. These abilities are grouped together because they allow an individual to operate according to the conceptual rules. The rules I have in mind here are the possession conditions of Peacocke's analysis.[83]

What distinguishes full-blooded concept-use from coincidental conformity with the rules is the subject's having a conception that they are part of a communal practice, and that they must try to march in step with that practice. This awareness is all that rule-following amounts to, no more, no less.[84] This awareness requires a certain higher-order ability to view one's own practice, i.e., concept is a personal level notion; but this need not rule out in principle a non-sentential analysis of cognition. What has stood in the way of connectionist attempts to model these higher cognitive phenomena is the lack of this global responsiveness. A level of complexity and sensitivity is needed to be a concept-user that such simple models could not hope to have, but they nevertheless suggest a picture of how concept-use develops. As an individual faces experience, they must create and deploy prototypes in order to bring about their goals, which could be as simple as food and comfort in an infant. Through experience the number of prototypes and their interactions increases. At some point the bare exercise of abilities comes to be seen as a part of a rule governed practice, and full-fledged concept-use begins. But this is not the end, as more and more complexity is worked into the system, and as the interrelations of a given vector prototype is increased. A child might learn that 'red' applies to objects of a certain colour, and apply it to everything that has that colour. Yet eventually it may come to realize that the concept, *red*, does not apply to objects that only appear red in certain environmental conditions, such as a red light, and that red objects are still red when there is no light shining upon them.

---

[82] J. McDowell, 'Functionalism and Anomalous Monism', in E. LePore and B. P. McLaughlin, eds., *Actions and Events: Perspectives on the Philosophy of Donald Davidson* (Blackwell, Oxford, 1985), pp. 387-398.

[83] C. Peacocke, *A Study of Concepts* (MIT Press, Cambridge, Mass., 1992).

[84] From some perspectives, namely those which see rule following as involving physically instantiated tokens of rules as a necessary part of the processing underlying concept-use, this is a minimal account. But from other perspectives this is a rich account because it involves conscious appreciation of rules.

The public system of rules to which individuals cede authority is an abstract ideal. No individual actually exemplifies it, and no individual could, because there is no complete public system. For one thing the public system of rules is not closed, it is an organic, developing structure. The concept *red* demonstrates this, for when the colour red was first referred to, in the mists of prehistory, it seems odd that this should fix the concepts use at all points, even in situations that might depend upon modern technology, such as artificial light. The rules are what the majority, or a body of experts decide upon. There is no greater authority to which one can apply. The objectivity of meaning springs from obedience to the rules. But here I have based an analysis of concepts upon a notion of meaning that I have not, as yet spelled out clearly.

Finally, it is important to note a criterion which any account of concepts must meet, and which the account given above does meet, namely developmental plausibility. We are not born with an innate conceptual scheme, *pace* Fodor. This constrains theories of concepts in a number of ways. First, they must be such as to be learnable in principle. This includes the fact that one should be able to establish that one has successfully grasped a concept. Any theory which leaves one in a position of having to make a guess that one has the concept in question is unacceptable, or at least it should be. Hence this constraint might be viewed as demanding that concepts be objective. Secondly, concepts should be learnable in practice; the time and processing limitations of human infants should not present an obstacle to concept acquisition.

Given these sorts of considerations I think that they point to another criterion of virtue for theories of concepts; if a theory can accommodate the gradual emergence of concepts in an individual this should be seen as a pragmatic boon. For any theory which trades on a light switch metaphor, i.e. a sudden transition into the charmed circle of concept-users, seems developmentally implausible. Rather a theory of concepts should explain how they can be the sorts of things that emerge gradually, so that the conceptual journey from neonate to adult is a continuous, if non-linear, one.

3.5 Summary

In this chapter I have attempted to draw out some philosophical conclusions from the discussion of connectionist and neural networks in the previous two chapters. I have used Deacon's analysis of symbolic reference as a basis for an account of how a parallel, neurally inspired, model of cognition might explain symbolic and linguistic thought. This involves the interactions of many different networks, with different computational roles. I have also made use of the idea of a suitably trained recurrent vector space, with grammatically constrained trajectories to explain how humans manage to obey the grammars of their languages. Thus the model that emerges is a hybrid, involving distributed networks, with sparsely labelled links to other networks, so that the results of a particular computation can go on to play an appropriate role in the system.

I have also suggested a philosophical analysis of meaning which is compatible with this empirical model, whose salient points rest upon the details of the physical implementation of the model. This makes good on my claim that there is a tight link between the findings of sciences interested in the mind and brain and the philosophy of mind.

4 Conclusions and Further Work

From the first reading of Paul Churchland's vector analysis of cognition I have been struck by its intuitive appeal and explanatory power. He has taken the idea of a multidimensional recurrent vector space with its powerful semantic metric and applied it to all aspects of human thought. Even science is encompassed: a scientist's understanding does not consist in a body of laws, but rather in having a well-configured weight space in a recurrent network. The space is well-configured because it allows experiential input to be mapped onto an appropriate prototype, which embodies the knowledge built up through previous experiences. These previous experiences allow him to understand how the present situation will unfold. Thus at the heart of the model are the processes of recognition and learned association. These notions crop up again in Deacon's account of symbolic thought, and I have tried to use these links to overcome some of the shortcomings in Churchland's model. In brief, I have argued that we must move beyond the analysis of single networks and onwards to the possible ways in which different kinds of network might interact. This might involve the introduction of new concepts and methods of analysis, but it does not vitiate the vital insights that emerge from investigation of connectionist networks, it is additive to that body of research. In this way we might come to a new understanding of how complex, conceptual behaviour might be produced by a brain whose mode of processing is fundamentally parallel, distributed, and recurrent. The processing is distributed in two ways: first, within individual networks as has been described at length, and second, in that a single task uses many different networks spread throughout the brain.

These are empirical speculations, but I have tried to sketch out how they might be integrated into a philosophical theory that encompasses meaning and understanding. I have also attempted to show how actual parallel brain processing could be compatible with a physical commitment regarding the instantiation of concepts. No doubt many of the details are wrong, but I believe that in broad outline something like the above will emerge over the next few decades of research of the mind. The key point that I hope emerges from this thesis, is that philosophy must be

responsive to the results of empirical speculation; it is not a domain apart, which cannot be touched by matters of fact.

Bibliography

G. T. M. ALTMANN, *The Ascent of Babel: An Exploration of Language, Mind, and Understanding* (Oxford University Press, Oxford, 1997).

P. BAK, *How Nature Works: The Science of Self-Organized Criticality* (Oxford University Press, Oxford, 1997).

J. L. BERMUDEZ, 'Peacocke's Argument Against the Autonomy of Nonconceptual Representational Content', *Mind and Language* (1994), 402-418.

G. G. BAYLIS, E. T. ROLLS, and C. M. LEONARD, 'Selectivity Between Faces in the Responses of a Population of Neurons in the Cortex in the Superior Temporal Sulcus of the Monkey', *Brain Research* 342, 91-102.

P. M. CHURCHLAND, 'Reduction, Qualia, and the Direct Introspection of Brain States', *Journal of Philosophy* 82 (1985), 8-28.

P. M. CHURCHLAND, *A Neurocomputational Perspective: The Nature of Mind and the Structure of Science* (MIT Press, Cambridge, Mass., 1989).

P. M. CHURCHLAND, *The Engine of Reason, the Seat of the Soul* (MIT Press, Cambridge, Mass., 1996).

P. M. CHURCHLAND, 'Learning and Conceptual Change', in A. Clark and P. Millican, eds., *Connectionism, Concepts, and Folk Psychology* (Clarendon Press, Oxford, 1996), pp. 7-43.

P. S. CHURCHLAND, *Neurophilosophy: Toward a Unified Science of the Mind-Brain* (MIT Press, Cambridge, Mass., 1986).

A. CLARK, 'Systematicity, Structured Representations and Cognitive Architecture: A Reply to Fodor and Pylyshyn, in T. Horgan and J. Tienson, eds., *Connectionism and the Philosophy of Mind* (Kluwer Academic Press, Boston, 1991), pp. 198-218.

A. CLARK, *Associative Engines: Connectionism, Concepts, and Representational Change* (MIT Press, Cambridge, Mass., 1993).

A. CLARK, *Being There: Putting Brain, Body, and World Together Again* (MIT Press, Cambridge, Mass., 1997).

A. CUSSINS, 'The Connectionist Construction of Concepts', in M. Boden, ed., *The Philosophy of Artificial Intelligence* (Oxford University Press, Oxford, 1990), pp. 368-441.

A. CUSSINS, 'Content, Embodiment and Objectivity: The Theory of Cognitive Trails', *Mind* 101, 651-88.

A. DAMASIO and H. DAMASIO, 'Cortical Systems for Retrieval of Concrete Knowledge: The Convergence Zone Framework', in C. Koch and J. Davis, ed., *Large-Scale Theories of the Brain* (MIT Press, Cambridge, Mass., 1994), pp. 61-74.

T. DEACON, *The Symbolic Species: The Co-evolution of Language and the Human Brain* (Allen Lane, The Penguin Press, London, 1997).

D. C. DENNETT, *Consciousness Explained* (Little, Brown, Boston, 1991).

F. DRETSKE, 'Misrepresentation', in R. Bogdan, ed., *Belief: Form, Content, and Function* (Clarendon Press, Oxford, 1986), pp. 17-36.

J. L. ELMAN, 'Representations and Structure in Connectionist Models', in G. T. Altmann, ed., *Cognitive Models of Speech Processing: Psycholinguistic and Computational Perspectives* (MIT Press, Cambridge, Mass., 1990), pp. 345-382.

J. L. ELMAN, E. A. BATES, M. H. JOHNSON, A. KARMILOFF-SMITH, D. PARISI, and K. PLUNKETT, *Rethinking Innateness: A Connectionist Perspective on Development* (MIT Press, Cambridge, Mass., 1997).

G. EVANS, *The Varieties of Reference* (Oxford University Press, Oxford, 1982).

J. FODOR and B. McLAUGHLIN, 'Connectionism and the Problem of Systematicity: Why Smolensky's Solution Doesn't Work', in C. Macdonald and G. Macdonald, eds., *Connectionism: Debates on Psychological Explanation* (Blackwell, Oxford, 1995), pp. 199-222.

J. FODOR and Z. PYLYSHYN, 'Connectionism and Cognitive Architecture: A Critical Analysis', in S. Pinker and M. Jacques, eds., *Connections and Symbols* (MIT Press, Cambridge, Mass., 1988), pp. 3-71.

P. GEACH, *Mental Acts: Their Content and their Objects* (Routledge and Kegan Paul, London, 1957).

J. J. GIBSON, *The Ecological Approach to Visual Perception* (Houghton Mifflin, Boston, 1979).

G. GILLETT, *Representation, Meaning, and Thought* (Oxford University Press, Oxford, 1992).

R. P. GORMAN and T. J. SEJNOWSKI, 'Learned Classification of Sonar Targets Using a Massively-Parallel Network.' *IEEE Transactions: Acoustics, Speech, and Signal Processing* (1988), 1135-1140.

R. F. HADLEY, 'Systematicity in Connectionist Language Learning', *Mind and Language* 9 (1994), 247-272.

D. H. HUBEL and T. N. WIESEL, 'Receptive Fields, Binocular Interaction, and Functional Architecture in the Cat's Visual Cortex', *Journal of Physiology* 160 (1962), 106-154.

M. I. JORDAN, 'Serial Order: A Parallel Distributed Processing Approach', Report 8604, Institute for Cognitive Science (University of California, San Diego, La Jolla, 1986).

S. M. KOSSLYN, *Image and Brain: The Resolution of the Imagery Debate* (MIT Press, Cambridge, Mass., 1994).

D. MARR, 'Simple Memory: A Theory for Archicortex', *Philosophical Transactions of The Royal Society of London, Series B* 262 (1971), 23-81.

D. MARR, *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information* (W. H. Freeman, San Francisco, 1982).

J. L. McCLELLAND, B. L. McNAUGHTON, and R. C. O'REILLY, 'Why There are Complementary Learning Systems in the Hippocampus and Neocortex: Insights from the Successes and Failures of Connectionist Models of Learning and Memory', *Psychological Review* 102 (1995), 419-457.

J. McDOWELL, 'Functionalism and Anomalous Monism', in E. LePore and B. P. McLaughlin, eds., *Actions and Events: Perspectives on the Philosophy of Donald Davidson* (Blackwell, Oxford, 1985), pp. 387-398.

J. McDOWELL, *Mind and World* (Harvard University Press, Cambridge, Mass., 1994).

P. J. G. NICHELLI, J. GRAFMAN, P. PIETRINI, K. CLARK, K. Y. LEE, and R. MILETICH 'Where the Brain Appreciates the Moral of a Story', *NeuroReport* 6 (1995), 2309-2313.

D. NORRIS, 'A Dynamic-Net Model of Human Speech Recognition', in G. T. M. Altmann, ed., *Cognitive Models of Speech Processing: Psycholinguistic and Computational Perspectives* (MIT Press, Cambridge, Mass., 1990), 87-104.

G. A. OJEMANN, 'Cortical Organization of Language', *Journal of Neuroscience* 11 (1991), 2281-2287.

G. A. OJEMANN and C. C. MATEER, 'Human Language Cortex: Localization of Memory, Syntax, and Sequential Motor-Phoneme Identification Systems', *Science* 205 (1979), 1401-1403.

C. PEACOCKE, *A Study of Concepts* (MIT Press, Cambridge, Mass., 1992).

W. PENFIELD and L. ROBERTS, *Speech and Brain Mechanisms* (Oxford University Press, London, 1959).

S. PINKER, *Learnability and Cognition: the Acquisition of Argument Structure* (MIT Press, Cambridge, Mass., 1989).

K. PLUNKETT and J. L. ELMAN, *Exercises in Rethinking Innateness* (MIT Press, Cambridge, Mass., 1997).

E. T. ROLLS, 'Parallel Distributed Processing in the Brain: Implications of the Functional Architecture of Neuronal Networks in the Hippocampus', in R. G. M. Morris, ed., *Parallel Distributed Processing: Implications for Psychology and Neurobiology* (Oxford University Press, Oxford, 1989).

E. T. ROLLS, 'Brain Mechanisms for Invariant Visual Recognition and Learning', *Behavioural Processes* 33 (1994), 113-138.

E. T. ROLLS and A. TREVES, *Neural Networks and Brain* (Oxford University Press, Oxford, 1998).

D. RUMBAUGH, ed., *Language Learning by a Chimpanzee: The Lana Project* (Academic Press, New York, 1977).

W. SELLARS, 'Empiricism and the Philosophy of Mind', in H. Feigl and M. Scriven, eds., *Minnesota Studies in the Philosophy of Science*, vol. 1 (University of Minnesota Press, Minneapolis, 1956), pp. 253-329.

R. M. SEYFARTH, D. L. CHENEY, and P. MARLER, 'Monkey Responses to Three Different Alarm Calls: Evidence for Predator Classification and Semantic Communication', *Science* 210 (1980), 801-3.

A. TREVES and E. T. ROLLS 'What Determines the Capacity of Autoassociative Memories in the Brain?', *Network* 2 (1991), 371-397.

D. C. VAN ESSEN, C. H. ANDERSON, and B. OLSHAUSEN, 'Dynamic Routing Strategies in Sensory, Motor, and Cognitive Processing', in C. Koch and J. Davis, eds., *Large-Scale Theories of the Brain* (MIT Press, Cambridge, Mass., 1994), pp. 271-299.

D. C. VAN ESSEN, B. OLSHAUSEN, C. H. ANDERSON, and J. L. GALLANT, 'Pattern Recognition, Attention, and Informational Bottlenecks in the Primate Visual System', *Proceedings of the SPIE Conference on Visual Information Processing: From Neurons to Chips* 1473 (1991), 17-28.

T. VAN GELDER 'What is the 'D' in PDP'? A Survey of the Concept of Distribution', in W. M. Ramsey, D. E. Rumelhart, and S. P. Stich, eds., *Philosophy and Connectionist Theory* (Lawrence Erlbaum, Hillsdale N.J., 1991), pp. 33-59.

L. WITTGENSTEIN, *Philosophical Investigations* (Blackwell, Oxford, 1953).

L. WITTGENSTEIN, *Zettel* (Blackwell, Oxford, 1967).