# Content and Its Vehicles in Connectionist Systems

## Nicholas Shea

**Abstract**

This paper advocates explicitness about the type of entity to be considered as content-bearing in connectionist systems; it makes a positive proposal about how vehicles of content should be individuated; and it deploys that proposal to argue in favour of representation in connectionist systems. The proposal is that the vehicles of content in some connectionist systems are clusters in the state space of a hidden layer. Attributing content to such vehicles is required to vindicate the standard explanation for some classificatory networks' ability to generalise to novel samples their correct classification of the samples on which they were trained.

Address for correspondence:
Somerville College, Oxford, OX2 6HD

Email address:
nicholas  shea (at) philosophy ox ac uk
(replace gaps with dots)

# Content and Its Vehicles in Connectionist Systems

**Contents**

An unnoticed misconception has had a pernicious effect in debates about connectionism. A tacit assumption has been at work about the entities to which content should be ascribed in connectionist systems. Lack of clarity about what they are taking the content-bearing entities to be is found on both sides of the ongoing debate about whether connectionist systems process internal representations. That battle, engaged most prominently between Jerry Fodor and Paul Churchland, can only be adjudicated relative to a specification of the vehicles to which putative contents are to be ascribed. Moreover, the assumption standardly made by connectionists, usually tacitly, about the individuation of vehicles of content serves to undermine their case for connectionist representation.

This paper advocates explicitness about the type of entity to be considered as content-bearing in connectionist systems; it makes a positive proposal about how vehicles of content should be individuated; and it deploys that proposal to argue in favour of representation in connectionist systems. To preview, the vehicles are clusters in the state space of a hidden layer; and attributing content to such vehicles is required for one kind of

explanation of the behaviour of some networks.

The paper is in seven sections. Section (1) brings to light some standard assumptions about the vehicles of content in connectionist systems and suggests how they can be relaxed. Section (2) proposes that clusters are the vehicles of content, and section (3) deploys clusters in a content-based explanation of how some networks manage to generalise their trained correct performance to new samples. Section (4) interprets Laakso & Cottrell's exciting empirical results about similarities between connectionist networks in the light of this proposal. Section (5) examines whether Fodor & Lepore's criticisms of state space semantics apply to the new approach. Section (6) mentions some further virtues of treating the vehicles of content in connectionist systems in this way, and section (7) lists some possible refinements.

## 1.    Standard Assumptions About Connectionist Representations

This paper is part of a larger project, which is to assess whether connectionist systems process internal representations. Can and should their operation be explained by attributing representational content to their internal states? My contribution is to formulate a clear proposal about the vehicles of content.[1] For the sake of clarity, the current paper considers only multi-layer feedforward classifier networks trained by supervised learning using a delta rule.[2] Section 7 below suggests how the proposal can be generalised.

The cognitive revolution rehabilitated internal representations. However, with the package came rejection of behaviourism's associative learning mechanisms and a requirement that representations combine compositionally. Connectionists who seek to explain their systems representationally need only subscribe to the first article of the canon. They aim to posit internal representations, but ones which are acquired by

---

[1]    I remain neutral about whether such contents qualify as 'meanings' in any sense.

[2]    Usher (2001) which uses probabilistic classification networks to analyse content, but in a different way: to get at the statistical relations which are relied on in the informational theory of content offered there. The treatment does not concern the contents found within hidden layers of classificatory networks.

associationist learning and which need not have compositional structure. Connectionists can still join the cognitivists in going beyond instrumentalism about representation – if they can show that their systems too process real internal vehicles of content.[3] Thus, representationalism requires, at a minimum, that a contentful explanation of the behaviour of a connectionist network be underpinned by a mechanism in the system which operates over representational vehicles to which those contents are ascribed.

Some may object to calling this 'representation', since it is sometimes required of representations that they enter into compositional structures, but the terminology is deliberate, since to assume that representations must be compositional is to beg an important question against the connectionist. The entities that vindicate the cognitive revolution's first and defining commitment – to an internal mechanism involving vehicles of content which are individuable non-semantically – deserve to be called representations. Typing representations as vehicles of content groups together different internal entities into classes that are importantly alike for internal processing, such they are all to be ascribed the same content: they are different realisations of the same vehicle of content.[4]

In a network that employs parallel distributed processing, each representation will be some pattern of activation distributed across a layer of the network. It is tempting to move from this truism to the tacit assumption that each pattern of distributed activation is a different representation, and thus that each can have a (slightly) different content.

---

[3]  There may be some organisms and systems that are only representers in the instrumental sense – they are interpretable from the intentional stance (Dennett 1987), though not in virtue of processing real internal representations. The issue here is whether a commitment to real internal representations, with its added explanatory purchase, can be defended for some connectionist systems.

[4]  As I use the terms here, representations are concrete particulars (with contents, in addition to non-semantic properties); and vehicles of content are also representations, but individuated in a particular way: according to the non-semantic properties which group different tokens together as being of the same representation type for the purpose of assigning content to them. The usage is the same with written language: the word 'John' is the vehicle of the content *John*, but the word is a type – a (non-semantic) typing of marks on the page, of which the following are tokens: John, John, JOHN. If a representation is picked out semantically (eg, my current singular thought about my brother), then this also picks out a concrete particular – a representation – but not as a vehicle of content.

However, that is to deny that a network's states can be grouped together into vehicles of content at all. We should not foreclose the possibility of specifying vehicle types that abstract away from details of their realisation. After all, vehicle types in a classical computer are multiply realised in the implementing mechanism, rather in the way that the ink marks 'dog', 'Dog' and 'DOG' are different physical realisations of the same word type. In searching for a non-semantic representation typing for connectionist systems, we should be open to the possibility that different patterns of distributed activation are of the same vehicle type.

The same point can be put in the language of state space. A state space is a useful way to think about patterns of activation across a layer of a network. State space is a notional high-dimensional space whose axes are constituted by the activation of nodes of the layer, so that any pattern of distributed activation corresponds to a point in the space. The standard assumption is that each point in state space is a different representational vehicle with a slightly different content. That is a mistake. A successful representational account of connectionist systems is likely to group the points of state space into vehicle types, such that different points are of the same vehicle type. At least, in searching for a non-semantic representation typing we should allow for that possibility. We should reject the standard tacit commitment, which implicitly excludes it, that vehicle types are maximally fine-grained, each point in state space being a different representational vehicle with the potential to have a unique content.

A second standard assumption is to take each individual node in a layer as representing something separately, so that the nodes which are active in a given distributed pattern of activation make individual (proportional) contributions to the content of the distributed representation. For example:

'Mental representations are taken to consist of "subsymbols" associated with each *node*, while "whole" representations are real-valued vectors in a high-dimensional property space.' (Eliasmith 2003, p. 2, italics added.)

'The activation of a given *unit* (in a given context) thus signals a semantic fact: but it

5

may be a fact that defies easy description using the words and phrases of daily language.' (Clark 2001, p. 67, italics added.)

The nodes are thought to represent complex, inexpressible 'microfeatures' of the presented stimuli. For example, Cottrell thinks that each node in one of his trained face recognition networks represents a 'holon': some complex property of human faces, which can often be visualised as a ghostly, face-like shadow.[5] This idea is both supported by, and lends support to, the foregoing assumption of maximally fine-grained vehicle individuation. The two assumptions together underwrite what I will call the microfeatural approach. The microfeatural approach encourages the view that, if connectionist systems represent, they do so in a way which is highly complex, with contents quite unlike those in everyday explanations. Their contents may even be ineffable. From here it is only a short step, via viewing connectionist networks as some kind of model of human brains, to eliminativism about the contents ascribed in everyday psychological explanations,[6] which some are happy to embrace.[7]

Yet there is no reason why the semantically relevant dimensions of a hidden layer state space should correspond to the activations of single nodes. The semantic dimensions may be independent of the axes defined by individual nodes. Thus, relaxing this second standard assumption opens up the possibility that hidden layers with different numbers of nodes could have the same number of semantic dimensions.

Finally, the microfeatural approach encourages the idea that hidden layer nodes represent even before training. After all, even when the network is assigned random or arbitrary weights, each node is differentially responsive to some complex, unspecifiable property of possible stimuli. However, the most exciting quality of connectionist models is their ability to develop entirely new representational resources. At best, they may give us an insight into how neural systems can develop from an untrained state which does not

---

[5]    Cottrell & Metcalfe (1991).

[6]    Ramsey, Stich & Garon (1990).

[7]    P.M. Churchland (1981); P. S. Churchland & T. J. Sejnowski (1989).

represent to a trained one which does. The microfeatural approach is apt to stifle that hope from the outset. We should reject it, and allow that it may be that vehicles of content only arise at all as a result of development, so that an untrained network may be incapable of representation.

TABLE 1. Standard assumptions about vehicles of content in connectionist systems, and alternatives

| | |
|---|---|
| (1) Each point in hidden layer state space has a (slightly) different content | (1') Different points in state space may fall into the same representational vehicle, and thereby have the same content |
| (2) Hidden units are the basic semantic dimensions of hidden layer state space (each representing some complex 'microfeature') | (2') Semantic dimensions in hidden layer state space may be independent of the axes defined by hidden layer nodes |
| (3) Points in hidden layer state space are content-bearing before training | (3') Points in hidden layer state space may not be content-bearing except as a result of development |

## 2. Clusters as Vehicles of Content

When a network is trained, the final values of connection weights between its nodes are exquisitely sensitive to the starting weights (which are often assigned randomly) and the order of presentation of training samples. It seems dizzyingly difficult to compare the different weight matrices which result. And there is no obvious way of comparing trained networks with different architectures, since their weight matrices have different dimensions. The microfeatural approach entrenches the view that these different matrices encode different solutions to the training task, which may be practically incommensurable. When that approach is rejected, the possibility arises that the mechanism of operation of a network can be described in a way which abstracts away from individual weight matrices and particular patterns of activation, such that it is shared by different networks trained on the same task. Such a description of a network's mechanism of operation would treat different patterns of activation as realisations of the same inner state, and so would be a good candidate for typing of vehicles of content.

It is striking that trained networks often manage to generalise: they can project their

correct performance to samples not encountered during training.  How so?  An optimistic hope is that there is a representational explanation of this ability.  This gives us a wish-list to guide our search for vehicles of content:

<u>Desiderata – typing the vehicles of content in a connectionist system should:-</u>

(i)     capture some underlying property of the network's mechanism of operation by which it performs its task;

(ii)    abstract away from individual weight matrices and particular patterns of activation;

(iii)   be such that it may be shared by different networks trained on the same task; and

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

Fodor claims that there is no way of describing a connectionist network that separates points in state space into different types:

'The "smallest" unit of connectionist representation for which a type/token relation is definable is a whole network.'  (Fodor 2000, p. 50)

The present paper takes Fodor's objection as a challenge – it aims to show that there is such a typing (and that it meets the foregoing desiderata for vehicles of content).

Fortunately, modelling work already has a resource which can be modified and redeployed to meet the desiderata – cluster analysis.  Cluster analysis is a family of techniques for mapping the distribution of activation points in state space.  In one version a dendogram is constructed which pairs each activation point with its nearest neighbour, then pairs these pairs with nearest pairs, and so on, producing a hierarchical tree. Dendograms show that in trained networks similar samples often produce similar patterns of activation in the hidden layer.  An example is Sejnowski & Rosenberg's (1987) NETtalk network, which was trained to map English text to phonetic representations of its

pronunciation. NETtalk's hidden layer makes a broad distinction between vowels and consonants, on the way to producing its fine-grained output classification into English phonemes (represented as distributions of articulatory features).[8]

Training tends to cause activation vectors to cluster together in hidden layer state space.[9] This is because the training task is to produce clusters in the state space of the output layer (correct responses to training samples fall into clusters, often degenerate clusters located on the axes constituting the space). The network must take points which are distributed throughout input layer state space and transform them into appropriate clusters in output layer state space. Forming a relevant intermediate clustering in hidden layer state space is a step towards achieving that goal. Indeed, modellers deploy various techniques to encourage hidden layer clustering.[10] The tools of cluster analysis can be understood as probing a network's representational structure – the modellers just don't realise that they are uncovering their systems' vehicles of content.

Making scatter plots of the receptivity of individual hidden layer nodes can also be understood as looking for clusters. Tight scatters are found if clusters happen to fall near an axis of the state space. Where two or more nodes are found to be receptive to the same set of samples (e.g. Dawson & Piercey 2001) it is especially clear that this reflects a single underlying cluster. Other techniques merely suggest that there is hidden layer clustering without individuating the clusters. For example, clusters cannot reliably be read off dendograms. To individuate clusters, the first step is to plot the distribution of points in
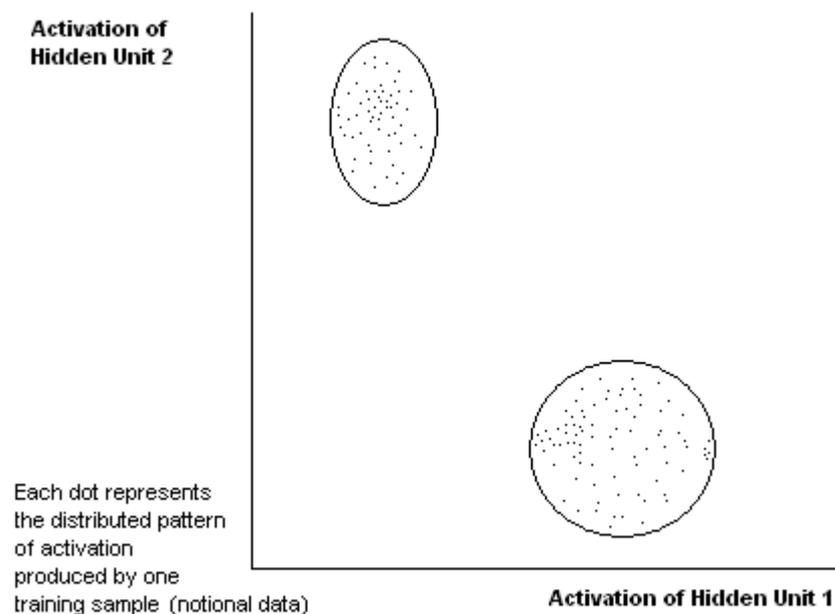
---

[8]  Of course, this is not a miracle. There must be information in the input-output mapping on which the network was trained that allows it to make the vowel-consonant distinction. In the version of NETtalk which codes outputs as phonological features, this information may be found in part in the fact that vowels are more alike one another in phonological features than are consonants. However, similar results are achieved when NETtalk's outputs are coded orthogonally, in which case the information needed to make the vowel-consonant discrimination must have been derived during training entirely from the inputs.

[9]  E.g. in Pollack's (1990) recursive auto-associative memory networks and Elman's (1990) simple recurrent networks, both described in Bechtel & Abrahamsen (2002), pp. 171-187.

[10]  E.g. 'skeletonization' in Mozer & Smolensky (1989); training in graded batches in Elman (1991); and extra-output learning in Dawson et al (2000).

the hidden layer state space of the trained network (each point corresponds to the activation produced by one of the samples to which the network has been trained to respond correctly). Regions of state space are then identified containing clusters of proximal points which are relatively distant from the other points in state space (relative to the overall volume filled by activation points produced by all training samples). There are various methods of determining how many clusters there are and where they lie. There are also important outstanding empirical issues about how best to individuate clusters in particular cases and whether to understand their boundaries as sharp or fuzzy. However, the important point is that a measure of proximity that is relative to the dimensions of a particular state space is sufficient to allow points to be grouped together based on such proximity.

FIGURE 1. Clusters in a state space with two hidden layer nodes



**Activation of Hidden Unit 2**

Each dot represents the distributed pattern of activation produced by one training sample (notional data)

**Activation of Hidden Unit 1**

My claim is that these clusters or regions in state space are vehicles of content. The system's internal mechanism can be described in terms of operations on these vehicles thus: a presented sample activates a hidden layer cluster, which in turn activates an output layer cluster. We will see in the next section how and why contents should be ascribed to such clusters.

Notice that different networks can have type-identical clusters – a cluster in the state

space of network A might be activated by the very same samples[11] as those which activate a particular cluster in network B. Indeed, clustering properties are independent of the number of hidden layer nodes,[12] so that state spaces with different numbers of axes may contain the same clusters. Thus, clustering properties can be shared by networks with different weight matrices and architectures. Networks might share some but not all clustering properties – a cluster in network A might correspond to another in network B activated by just the same samples without there being a correspondence between the other clusters in the two networks. So sharing some clusters does not guarantee that the networks will have the same overall set of vehicles of content.

Are clusters causally efficacious? The worry is that the activation of a cluster is always realised by the activation of a particular pattern of distributed activation. A complete description of the operation of a network can be given at this lower level, in terms of connection weights and individual patterns of activation. This worry is, in fact, just a particular guise of the standard problem of the causal efficacy of the properties found in the special sciences. It arises for vehicles of content in a classical computer, too, since classical syntax is also realised by lower-level processes. The syntax of a high level programming language is realised by the syntax of lower-level languages, sometimes proceeding via a series of levels until primitive computations are arrived at which can be implemented in physical components like transistors (whose operation is explained by molecular properties of the semiconducting crystals which realise those components). Indeed, it should arise for any adequate typing of vehicles of content in any computational system, since the vehicle types should generalise over some class of lower level causal / mechanistic entities. It is an important metaphysical question whether any such special science property is causal – perhaps overdetermination can be dealt with and causation

---

[11] Throughout I use 'samples' for the real world entities, like colour chips, that are coded into inputs on which the networks are trained. Thus, the same sample will be encoded into different input vectors for the training of different networks. Points in a cluster are compared via samples, not input vectors.

[12] Subject to the constraint that the number of independent dimensions in cluster space obviously cannot exceed the number of nodes in a layer.

exists at multiple levels, perhaps they merely cause in virtue of being realised, etc. – but the causal efficacy of connectionist clusters is not a special case. For present purposes it is enough that clusters are just as explanatory, or as causal-explanatory, as the properties of chemistry, electronics or biology; indeed, they have the same metaphysical status as vehicles of content in classical computers. Most notably, these metaphysical questions arise for mental properties in general, when considered naturalistically. It might even be thought a merit of my proposed vehicles of content that, on this issue at least, they have the same metaphysical status as mental properties.

Where modellers have only tacitly subscribed to the microfeatural approach, it is easy to reconceive their explanations in terms of clusters. By contrast, Andy Clark has been much more careful, and he explicitly endorses the microfeatural approach to understanding static networks (1993, 1996; and 2001 quoted above). Even so, when it comes to dynamic connectionist networks, Clark abandons the microfeatural idea, which is hard to make any sense of in the dynamic context. Instead, he allows that dynamic analysis of a network – finding attractor basins or principal component processes that account for its behaviour over time – might be uncovering temporally extended physical processes that are the vehicles of representational content in such systems (Clark 2001, p. 135). The complexity of the dynamic case pushes Clark away from seeing such systems in terms of microfeatures. I would argue that the basic unworkability of the microfeatural idea is just as good a reason for abandoning it for static networks. If basins of attraction for dynamic processes are the vehicles of content in dynamic networks, that strongly suggests that their analogue, clusters, are the vehicles of content in static networks.

## 3.    Contentful Explanation

So we have seen that a trained network's mechanism of operation can often be described in terms of clusters. Are these vehicles of content? The acid test is whether content should be ascribed to the clusters. I will argue that it should if clusters are to be invoked, as is common empirical practice, in an explanation of a network's ability to generalise its correct performance to new samples; thus that clusters are vehicles of content to the

extent that they form the basis of generalisation.

It is one of the remarkable features of connectionist networks that they can often correctly classify new samples which differ in their input encoding from anything encountered during training. Modellers standardly point to clustering to explain this phenomenon. It is observed that the activation produced by correctly classified novel samples often falls into existing clusters in a hidden layer,[13] and the ability to project correct behaviour to new samples is explained by the proximity to existing points in hidden layer state space of the activation produced by the new samples.[14] Conversely, where the hidden layer state space fails to differentiate into clusters, that fact is offered to explain why the network failed to project its correct responses to new samples.[15]

Describing the mechanism in terms of clusters can show why the network behaves as it does with new inputs. Considered as individual patterns of activation, the new inputs are not the same as anything that the network has encountered during training (worse, a new sample will generally not be linearly separable, in input layer state space, from points representing training samples mapped to entirely different output classifications). But from the point of view of clusters, the new samples produce activation in the same hidden layer clusters as samples in the training set. When the network is described as transforming samples into hidden layer clusters and onwards into output layer clusters, then it is apparent that the new samples *are* being treated in the same way as some of the samples in the training set.

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

---

[13]    Lehky and Sejnowski (1987) & (1988), Hinton (1989), Elman (1991), Dawson & Piercey (2001).

[14]    Churchland & Sejnowski (1992), p. 169.

[15]    Clark (1993), pp. 132-135, Elman (1991).

of explanation: the new samples have some property in virtue of which they fall into existing hidden layer clusters, and so cause the network to produce correct responses at the output layer.

That kind of explanation cannot just advert to patterns of activation at the input layer, since new samples differ in their input encodings from anything in the training set. So it is obliged to advert to properties of the samples themselves. The explanation is that the network is able to keep track of some property that is common between a new sample and some of the samples in training set, and that is relevant to the output classification. It does so by means of hidden layer clusters. That is to say, by activating an existing hidden layer cluster the network represents that the new sample has the property common to the training samples in the cluster. This correct intermediate classification is part of the process by which the network makes a correct output classification. Conversely, if a new sample fails to activate an existing hidden layer cluster, or if it activates a cluster of training samples with which it does not share a relevant property, the network will usually fail to make the correct output classification. In that case, misrepresentation at the hidden layer helps account for misrepresentation at the output layer.[16] So, are modellers right to point to hidden layer clusters to explain how a network manages correctly to classify a novel sample? Yes, provided their explanation is understood as attributing content to those clusters.

In short, the ability to project to new samples is explained by the fact that hidden layer clusters represent properties of the samples which are relevant to the output classification. A more limited explanation is also available: the network correctly classifies samples (new and old) in the relevant domain (e.g. colour samples) by output properties $O_1$, …, $O_n$ (e.g. red, blue and green) because that is what it was trained to do. It is a historical explanation in terms of facts about the network's development. The explanation we have been considering goes further, and gives an ahistorical account of how, after learning is completed, the network correctly classifies a new sample. It relies, not on facts

---

[16]  Of course, correct classification at the hidden layer does not necessitate correct performance at the output layer, it just makes it much more likely; so the explanation offered here is abductive and defeasible.

about the network's developmental history, but on facts about its internal structure and the dispositions of those structures in relation to features of the network's environment. The explanation points to an intermediate processing stage which is sensitive to a relevant external property but which is, at the same time, part of the internal mechanism by which the system arrives at its output classification.

The argument for viewing clusters as vehicles of content is based on giving an explanation for generalisation. Some networks fail to generalise, for example by having too many nodes in their hidden layer and so treating each input 'near-individually'. In such cases, even if there happen to be clusters in the network's hidden layer state space (perhaps duplicating clusters in input layer state space), they will not be bearers of content.

The properties tracked by hidden layer clusters can relate in various ways to the properties the network is trained to represent at the output layer. They may be more general than, the same as, more specific than, or orthogonal to the output properties. We saw above that NETtalk divides samples into vowels and consonants on the way to making phonetic classifications at the output layer. Hinton's (1989) network trained to keep track of family relationships between individuals was sensitive, in its internal structure, to features like age and nationality which were not given as training primitives. Elman's (1990) simple recurrent networks for predicting the next word in a linguistic corpus are trained merely with series of binary encodings of words. Cluster analysis of a trained network showed that the hidden layer had organised the words into grammatical and semantic categories: nouns vs. verbs, within nouns into animate vs. inanimate nouns, and within animate nouns into words for humans vs. animals. A final example is Pollack's (1990) recursive auto-associative network trained on sets of syntactic phrase structure trees. After training, verb phrases formed one cluster in the hidden layer state space, and prepositional phrases another – a level of generality not given explicitly in the training data.

The proposal is not to endow networks with original intentionality. Hidden layer clusters are only contentful in virtue of the contents ascribed to outputs. The modeller takes the outputs to represent some properties $O_1, …, O_n$ and trains the network to be good

at classifying by these properties. Causal, informational and teleological theories of content would all ascribe content to the outputs on this basis, and even non-naturalistic approaches can allow that outputs have contents which derive from the intentions of the modeller (in the same way that words in a public language derive content from the intentions of speakers and hearers). The content of hidden layer clusters stems from their relevance as an intermediate stage in making this contentful output classification. In fact, even the individuation of vehicles of content (section (2) above) assumes that the system's response to samples can be judged as correct or incorrect: recall that hidden layer clusters are to be individuated by considering only samples *correctly* classified as a result of training (and plotting the activation they produce in hidden layer state space). Although hidden layer content is merely derivative, it is an important step towards representationalism about connectionist systems if we are able to understand the operation of hidden layers in contentful terms, given contents at the output layer.

So, we have the following sufficient condition for ascribing content to internal states of a connectionist system.


<u>Jointly sufficient conditions for attributing content to patterns of activation of a hidden layer of a connectionist network:-</u>

(a)    the network is able correctly to classify some set of samples which differ (i.e., differ in their input encoding) from those in the training set ('new samples');

(b)    the new samples fall into hidden layer clusters formed by samples in the training set;

(c)    each new sample shares a property with the training samples in its cluster; and

(d)    that property is relevant to the classificatory task.


The extent to which connectionist networks satisfy these conditions remains an empirical question. We saw above that there are good reasons to think that at least some do. Clearly, it is an idealization. In practice some points may fall outside any cluster, and some points in a cluster may not share a relevant property with the majority of their neighbours.

That is to be expected. When real systems approximate to the idealized model, the model can be used to explain the behaviour of the real system.

Since samples activating the same cluster are treated by the hidden layer as similar, a network should be seen as treating samples as different when they activate different clusters in a given layer. Thus, the content ascribed to a cluster should be distinctive of samples in that cluster, as compared with other clusters in the same layer. So the content to be ascribed to make good the foregoing explanation of correct classification of new samples is as follows:

Content of a cluster

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

This is not intended to be constitutive of content. Rather, it describes the contents that should be ascribed, but does not attempt to capture the factors in virtue of which clusters have those contents. Thus, it furnishes a useful constraint on any metaphysical theory of content – that, when applied to connectionist systems, the theory should deliver these contents.

Although I have expressed the content of a cluster in the form *the presented sample has property P,* the cluster has no constituent structure. These are non-conceptual contents (according to one common understanding of that term). English forces us to use a phrase with subject-predicate structure to describe a complete propositional content which, for the system, is realised by a single state without such structure. The system is doing something simple, like feature placing. It can represent properties of the currently-present sample but cannot represent properties of objects presented in any other way. For these simple systems, the connectionist network should not be thought of as representing propositional constituents, and its representations do not enter into compositional structures.

Since the properties tracked by a hidden layer must be causally or constitutively relevant to the output task, they must be natural properties. This is a notoriously difficult distinction to draw, but it is motivated by many considerations, and the distinction is needed in many fields; it is not peculiar to connectionist content. Roughly, the idea is that natural properties must figure in natural laws – in this case laws (causal or constitutive) that relate the properties represented at the output layer with those kept track of by the hidden layer. In particular, arbitrary disjunctions of properties will not do.

Will ascription of content to hidden layer clusters encourage eliminativism, as with the microfeatural approach (when connectionist systems are taken to model aspects of human psychology)? The four examples of hidden layer clustering mentioned above tracked properties which are familiar in folk psychology (vowel vs. consonant, nouns vs. verbs, etc.). Further empirical investigation is needed to be sure that this is true in general, but there is no obvious pressure towards complex or inexpressible contents. It seems likely that, when networks are trained to represent familiar properties at their outputs, the properties tracked by hidden layer clusters as a means to making such output classifications will usually be familiar as well. Contrast the microfeatural approach, where fine-grained vehicle typing leads to extremely complex contents, and where taking individual hidden layer nodes as the representational primitives involves attributing to them extremely unfamiliar contents. Those pressures towards eliminativism are mistaken according to the current proposal. The microfeatural approach gets the wrong contents because it individuates vehicles of content in the wrong way.

In sum, there is a good reason to attribute content to hidden layer clusters. They also satisfy our desiderata for individuation of vehicles of content (see §2 above): They describe a network's mechanism of operation in a way which abstracts away from individual weight matrices and particular patterns of activation such that the same operations may be found in different networks trained on the same task. Clusters also fulfil the optimistic ambition that representations might help to explain networks' mysterious ability to project correct classificatory practice to novel samples. Taken together, this amounts to a compelling case that clusters are vehicles of content in connectionist systems.

## 4.    Laakso & Cottrell's Results

The microfeatural approach takes it that individual hidden layer nodes are the basic vehicles of content.    Distributed patterns of activation are composed of these representational primitives, and inherit their content from them.  I have been arguing for a way of individuating vehicles according to which only the distribution of points in hidden layer state space is relevant, irrespective of their disposition in relation to the axes defined by hidden layer nodes.  Important empirical work by Laakso & Cottrell (2000) makes it strikingly clear that this is the right strategy.
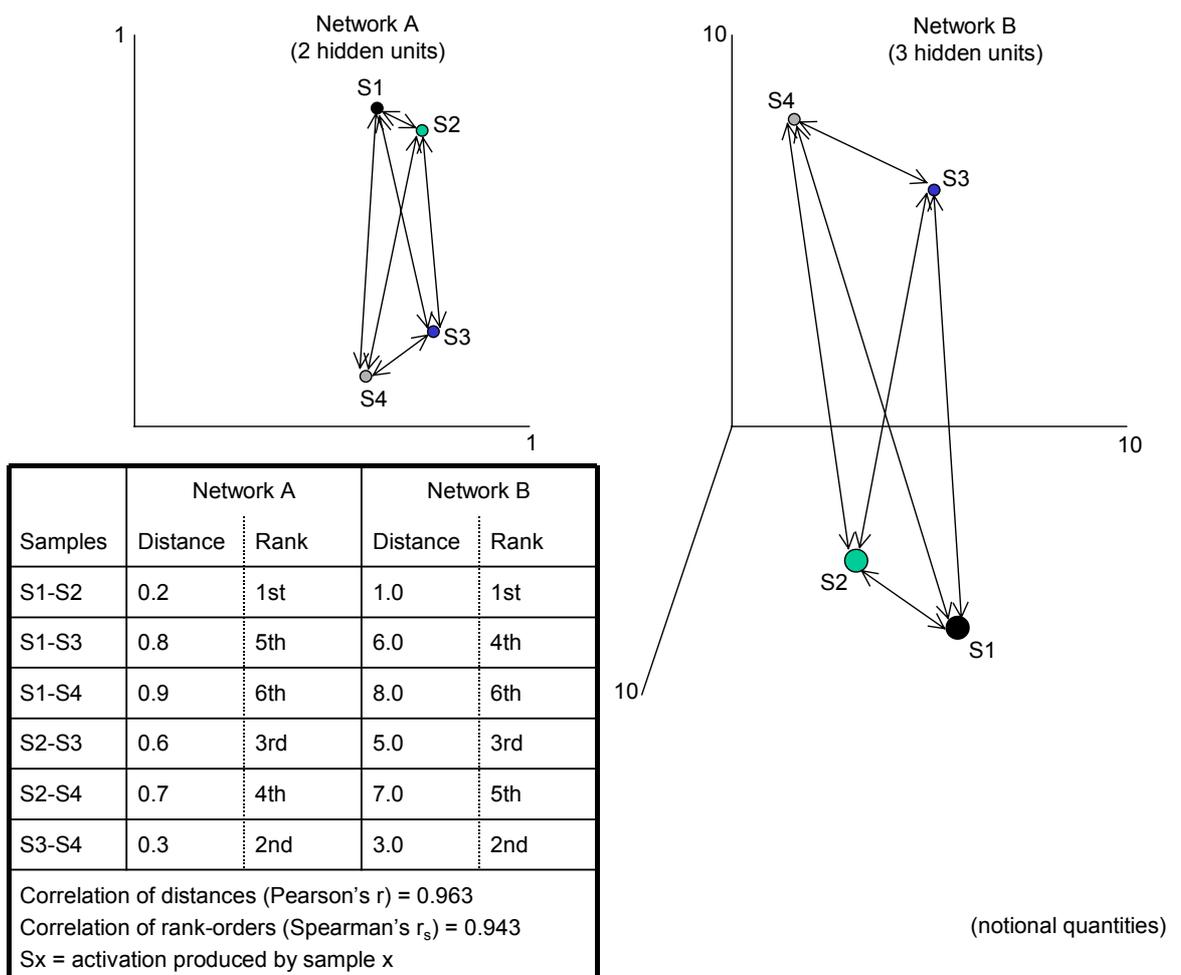
Laakso & Cottrell trained a series of networks with different architectures to do colour classification.  All were static three-layer feedforward networks learning by a delta rule with backpropagation of error.  They differed in the way colour samples were coded as inputs (ranging between 3 and 96 input nodes) and in the number of nodes in the hidden layer (1 to 10).  Laakso & Cottrell aimed to compare the arrangement of activation points in the respective state spaces of the different networks.  Remarkably, when trained to classify the same samples according to the same output classification, networks with different architectures arrived at very similar arrangements of activation points in hidden layer state space, provided they had three of more units in the hidden layer.

To show this, Laakso & Cottrell first constructed dendograms which indicate that their trained networks do show clustering.  Then they compared pairs of networks to see whether they clustered the samples in the same way.  Specifically, they tested whether samples that produced proximal activation in the hidden layer state space of one network also produced proximal activation in the other and, conversely, whether distal samples in one network were also distal in the other.  They did so by calculating the distances between pairs of points within each state space.  These distances can be thought of as a series of ordered pairs:

<distance between activation produced by sample m and sample n in network A,

distance between activation produced by sample m and sample n in network B>

To test how similar the state space of network A is to the state space of network B, Laakso & Cottrell measured the statistical correlation between the elements in these ordered pairs. This measures whether distances in network A which are small relative to the other distances in the hidden layer state space of network A correspond to distances in network B which are small relative to the others in network B, etc. The test is independent of the absolute size of either state space. In work reported by Churchland (1998), Laakso & Cottrell also compared the rank-orders of distances in the two networks, with the same results − networks trained on the same problem tended to have similar arrangements of activation points in hidden layer state space.

FIGURE 2.  Comparing the arrangement of points in two state spaces



| Samples | Network A | | Network B | |
|---|---|---|---|---|
| | Distance | Rank | Distance | Rank |
| S1-S2 | 0.2 | 1st | 1.0 | 1st |
| S1-S3 | 0.8 | 5th | 6.0 | 4th |
| S1-S4 | 0.9 | 6th | 8.0 | 6th |
| S2-S3 | 0.6 | 3rd | 5.0 | 3rd |
| S2-S4 | 0.7 | 4th | 7.0 | 5th |
| S3-S4 | 0.3 | 2nd | 3.0 | 2nd |

Correlation of distances (Pearson's r) = 0.963
Correlation of rank-orders (Spearman's $r_s$) = 0.943
Sx = activation produced by sample x

(notional quantities)

The beauty of these kind of tests is that they are applicable between networks with different architectures. Laakso & Cottrell found that even networks with different numbers of input nodes, or different numbers of hidden nodes, arranged activation points similarly in state space. This vindicates the intuition that networks with quite different weight matrices and even different architectures might, as a result of training on the same samples, solve a problem in the same way. It also shows that the kind of inter-network similarity in clustering properties suggested in my discussion of connectionist representation is found in real systems, and so lends further empirical support to the claim that clusters are vehicles of content.

Churchland (1998) welcomed Laakso & Cottrell's results (reporting them before their publication in 2000). He embraced the idea that state spaces constituted by different numbers of hidden nodes might nevertheless have points arranged in the same way, and thus that the semantically-relevant dimensions in state space are independent of the axes defined by individual nodes. In the process he gave up his earlier claim that individual hidden layer nodes represent complex microfeatures of stimuli presented to the network. However, he took the Laakso & Cottrell test to be a direct measure of content similarity. In my view, the test is too strong as a measure of content similarity. Networks which show clustering and whose inter-point distances are highly correlated will indeed have the same contents ascribed to all their corresponding clusters. But recall that on my approach it is possible for some clusters to share contents while others do not. Taking Laakso & Cottrell's correlation test as a direct measure of content similarity, as Churchland does, rules out this possibility. Churchland takes each individual point to be a different vehicle of content and is left with the whole state space as the only unit of comparison. Section (3) above shows why content should be ascribed to individual clusters. As a result, the contents of particular clusters can be compared. Then it can be that, for some clusters in network A there is a cluster with the same content in network B, but not for others. Thus, it does not follow from there being one or more corresponding pairs of clusters which have the same content in both networks that the networks have the same overall set of vehicles of content. Also, clusters with the same contents may be differently arranged in the state

spaces of two networks.[17]   In either case, Laakso & Cottrell's test would show a low correlation even though some or all of the clusters in the two networks do indeed have the same contents.   In short, Laakso & Cottrell's work establishes the important empirical result that, in practice, there are networks whose hidden layer clusters have the same contents.   But the arguments in the present paper show that a high correlation on the Laakso & Cottrell test is not necessary for some or all clusters to have the same content in the two networks under comparison.

## 5.    Fodor & Lepore's Criticism of State Space Semantics

Fodor & Lepore (1999) assemble various criticisms of Churchland (1998).   Churchland (in progress) has answers in relation to his account.   In this section I will examine whether Fodor & Lepore's criticisms run against the version of state space semantics which I have advocated above.   Fodor & Lepore have two main lines of objection.   The first is to oppose similarity-based accounts of content and the consequences of individuating contents holistically.   Secondly, they argue that Laakso & Cottrell's test could not be a measure of similarity in any event.

Fodor & Lepore rightly point out that an account of content similarity would be incoherent if there were no such thing as content identity.   However, Churchland need not concede that content identity is impossible. In his view:

'A point in activation space acquires a specific semantic content not as a function of its position relative to the constituting *axes* of that space, but rather as a function of (1) its spatial position relative to all of the *other contentful points* within that space; and (2) its causal relations to stable and objective *macrofeatures of the external environment*.'

(Churchland 1998, p. 8, his italics)

---

[17]    Calvo Garzón (2003) uses this as an objection to Laakso & Cottrell's test as a measure of content similarity, and rejects state space semantics as a result.

Churchland does not say exactly what the relevant function is, but however this framework is filled out networks with the same architecture and weight matrix embedded in the same problem will have the same contents. The objection can only be that identity is rarely realised in practice, not that it is impossible in principle. Churchland accepts this consequence, and therefore accepts the need for a robust notion of concept similarity.

Fodor & Lepore object to the idea that Churchland's content similarity can do the necessary explanatory work. To start with, from the quotation above it looks as if regress threatens. The identity of a particular pattern of activation depends upon its relation to all other contentful points in the same state space, the identity of each of which depends upon their relations to all other points (including the original one), etc. It follows that individual points in two state spaces cannot be compared, and that the only possible comparison is between the overall state spaces. Churchland appears to agree. His measure of content similarity only applies between whole state spaces. It is a standing property. But we want to allow for occurrent states: when the hidden layer of a network is differently activated on different occasions the network represents different things. Networks with very similar arrangements of points (possible activations) in state space nevertheless may be in different occurrent states (current activation) from one another. Relying only on a global measure of content similarity no such distinction can be drawn.

Thus, Churchland's proposal threatens a destructive holism in the individuation of representations.[18] Tiffany (1999) argues that Churchland's theory should avoid this difficulty by presupposing a theory of content. Churchland could then be read as proposing a way of comparing arrangements of contentful points in state space which is parasitic on a pre-existing assignment of content to those points. Tiffany thinks that state space semantics is best seen just as a way of individuating the representational vehicles. While I agree that this debate calls for clarity about the individuation of vehicles of content, I don't see how a pre-existing theory of content would help, because it can only assign contents once the bearers of those contents have been identified.

---

[18]    Fodor (2000), p. 52.

According to my proposal, by contrast, the vehicles of content (clusters) are individuated without presupposing anything about their contents. All that is needed is to plot the activation points in state space produced by the training samples. This allows clusters to be individuated without any commitment about what is represented by activation in the hidden layer. There is no threat of regress or holism in the individuation of vehicles of content.

Nor does holism about content follow from my proposal. A cluster's relations to other clusters do not determine its content. It is perfectly possible for networks to have clusters with the same contents arranged differently in hidden layer state space; and for two state spaces to be such that some contents are the same in both state spaces while others are completely different. There is a constraint that different clusters in the same layer should be assigned different contents, but this does not depend upon the arrangement of clusters in state space, nor does it entail content holism. Thus, Fodor & Lepore's objections to reliance on content similarity in general (for example, in translation) are not relevant to my proposal. (Furthermore, a robust measure of content similarity for individual activations may well be able to overcome Fodor & Lepore's general concerns about similarity-based semantics.)

Although it differs in important respects from Churchland (1998), it is worth emphasising that my proposal is inspired by his work and the results of Laakso & Cottrell on which he relies.[19] Indeed, it can be seen as a version of state space semantics, filling out the details of the function intimated in the quotation above, but with the caveat that relations between points in state space (Churchland's (1)) are relevant only to individuation of vehicles of content, whereas causal relations to external features ((2) above) are relevant to content ascription.[20]

Fodor & Lepore have a second line of objection. They argue that state spaces with

---

[19]   My proposal differs in substantially the same way from O'Brien & Opie (2001), which is another working-out of state space semantics in the light of Laakso & Cottrell (2000).

[20]   Notice that, on my proposal, the relevant external features include both properties of the input stimuli which cause activation of a cluster and the property represented by the output cluster which activation of that hidden layer cluster itself causes.

different numbers of nodes cannot have activations with the same content. Thus, any account of content similarity (Churchland's) or identity (mine) on which they do must be mistaken. The argument is that, if semantic properties depend upon positions in state space, then when state spaces have different dimensionality, points in those spaces will have different contents. For example, the point in 3-space which means *heavy and hard and black* (putative ROCK) along three semantic dimensions is quite different in content from the point in 4-space which is *heavy and hard and black and animate* (not a candidate for ROCK). The problem with Fodor & Lepore's objection is that it obviously assumes that state spaces with different numbers of hidden nodes have different numbers of *semantic* dimensions. That does indeed follow from the old microfeatural idea that individual nodes are the representational primitives. Churchland (1998) wisely abandons this idea.[21] Since he does not state explicitly that he has changed his mind ('I stand by those earlier responses …'[22]), it is understandable that Fodor & Lepore should predicate their objections on the standard microfeatural assumption found in Churchland's earlier work. This is where Laakso & Cottrell's results have been so important in changing the terms of the debate. They show that networks with different numbers of hidden layer nodes may nevertheless have the same semantic dimensions.

According to my proposal, too, semantic dimensions are independent of the axes of state space corresponding to individual nodes. However, since content is assigned to a cluster independently, and does not depend upon its relations to other clusters, even state spaces with different semantic dimensions may have some clusters with the same content in both networks.

Fodor & Lepore continue. Since they have demonstrated, they think, that Laakso & Cottrell's test cannot measure semantic similarity, they consider whether it might be merely a test for 'neural' similarity (i.e., similarity in the system's mechanism of operation). Recycling the previous argument, they contend that activation patterns in state spaces with different numbers of hidden nodes are incommensurable – they cannot be

---

[21]    Tiffany (1999) also interprets Churchland (1998) this way.

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

of the same neural type.[23]   They then suggest that it is a confusion, in any event, to suppose that individuating neural states is a way of individuating semantic states, since the two are found at different levels of description.  It is clear that Fodor & Lepore's objection misses the mark, however, once it is established that clusters should be the vehicles of content under comparison.  The microfeatural approach does indeed take a fine-grained mechanistic level of description to be the same as the semantic level.  By contrast, clusters provide a way of grouping different states of the mechanism into types, so that the vehicles of content generalise over many states of the mechanism.  Nevertheless, *pace* Fodor & Lepore, this vehicle typing is also a mechanistic level of description, since it is a way of describing the system's mechanism of operation.  An appropriately typed 'neural' level is also the semantic level.  It is not a confusion to take a neural state typing to be a typing of semantic states.  The semantic space just is a way of carving up neural space (when neural space is partitioned appropriately).

Fodor has a standard rejoinder to claims that the semantic level is a higher-level description of the operation of a connectionist network.  He says that the connectionist network is then just a way of implementing classical computation.[24]  That objection clearly does not count against clusters as vehicles of content.  They are indeed bearers of content, but in other respects they have very different properties to representations in classical computational systems, so that it is clear that connectionism is a quite different way of modelling psychology.  Clusters are processed differently from classical symbols.  They have no compositional structure.  They show different patterns of breakdown.  And they arise quite differently in development – as we shall see in the next section – which gives connectionism the potential to revolutionise our understanding of how the mind could be a computational system.

---

[23]   Fodor (2000), p. 51 states that they cannot; Fodor & Lepore (1999), pp. 399-400 suggests that they cannot, especially if the neural similarity is also the basis for content similarity.

[24]   E.g. Fodor & Pylyshyn (1988).

## 6.     Further Virtues of the Proposal

Perhaps the most striking virtue of my account of connectionist representation is the way it models representational development.  Theories of content standardly fail to engage with the mechanisms of representational development.  They just take the prior development of vehicles of content for granted – they assume a supply of pre-existing entities that can then be put into the internal and mind-world relations needed in order for them to be contentful.[25]  This assumption pushes Fodor towards his implausibly strong concept nativism (Fodor 1998).  By contrast, my proposal explains how a system develops new vehicles of content: training causes hidden layer clustering.  It offers a non-semantic (non-cognitive) explanation of the development of entirely new representational items.[26]  A connectionist system has no vehicles of content when it is assigned arbitrary connection weights before training begins.  It is only as a result of training that inputs are clustered together in the hidden layer.  Thus, only after training is a system representational at all, since only then does it have any vehicles of content.  That is a virtuous consequence.

Compare the microfeatural idea.  According to that approach, patterns of activation are vehicles of content whether or not any training has taken place.  They are compounds of the activations of individual nodes.  And even when connection weights are set at arbitrarily, some complex disjunctive microfeatures can be ascribed to each individual

---

[25]     E.g. Laurence & Margolis (2002), p. 42 (even as they criticise Fodor's strong concept nativism).

[26]     Unlike Tiffany (1999), as discussed in the previous section.  See Rupert (2001) for an argument that the development of new representational primitives must be explained in non-cognitive (ie, non-semantic) terms.  Rupert was responding to Cummins' general argument against representational theories of mind (Cummins 1997): that the development of new terms in the language of thought can never be explained non-semantically (since appealing to learning is always to explain development in terms of cognitive causes).  Rupert suggests that connectionists can answer this challenge by explaining the development of new representations non-cognitively and suggests that regions of state space may be the vehicles of content (although he talks in terms of concepts, rather than non-conceptual representations).  My proposal about the vehicles of content in connectionist systems vindicates that suggestion, and shows exactly how the connectionist can meet Cummins' challenge by explaining the development of new vehicles of content non-semantically.

node, on the basis of the samples that would cause it to be activated. I take it to be almost a *reductio* of the microfeatural idea that it can ascribe contentful states to an untrained connectionist network which consists simply of some architecture of nodes connected by arbitrary weights. We have no reason at all to think that the states of such a system are contentful. Yet the rationale for thinking of the hidden layer nodes as each encoding some complex feature of all the samples by which it is causally activated applies equally to the untrained network.

The syntax of a classical computer is also built-in. A common concern about modelling cognition on classical computation is that the primitive representations must all be present at the outset. The system can only develop by forming new complex representations out of these pre-existing components. That is a substantial limit. By contrast, connectionism furnishes a model of how new primitive representations develop – by the system learning to 'representationally redescribe', in its hidden layer, the categories given to it as outputs, or the problem in which it is embedded (Karmiloff-Smith 1994). Once clusters are seen to be vehicles of content, representational redescription can be explained: training gives rise to new representational vehicles (clusters) in hidden layer state space. The approach also shows how representation can arise out of training on a realistic action-based task.

The clustering approach to content and its vehicles in connectionist systems has many other virtues. There is scope here only to list them, without further explanation. It can explain the prototypicality effects exhibited by many trained networks. It has the merit of assigning a role to both inputs to and outputs from a system in ascribing content. Furthermore, to the extent that the existence and content of clusters idealizes away from the messy details of real systems, it shows why processing described at the semantic level may only satisfy 'soft' constraints.[27]

---

[27]    In the terminology of Smolensky (1988), but for a different reason.

## 7.    Possible Refinements

For the sake of clarity, I have only presented the simplest version of the proposal that the vehicles of content in connectionist systems are clusters or regions of state space.  I should mention several possible refinements.

The proposal is naturally extended to recurrent networks, where the vehicles of content are basins or regions of attraction for the dynamic processes which the system undergoes.  In that case, the vehicles are not be confined to a particular layer, but will be realised by the entire network, nor are they temporally confined since they are not states but temporally-extended processes.  The proposal also applies to networks which develop with Hebbian (or any localist) learning rules and unsupervised learning algorithms.  For example, the processing of trained pattern association, auto-association and competitive networks[28] can all be described in terms of clusters.  Furthermore, by attaching further networks to a hidden layer in which clustering has occurred, clusters could be taken as input for further kinds of onward processing; a simple example being to map the clusters to individual neurons using a competitive network.

The proposal can be extended to the individuation of clusters of clusters, and thus form the basis for an explanation of representational nesting that does not reply upon compositional structure or explicit knowledge representations.  It is also consistent with the analysis of hidden layer state space in terms of principal components.  Where some clusters in a layer are the result of the simple combination of others, then those principal components can be viewed as the primitive representational entities.[29]  The individuation of clusters can also be modified slightly to take account of the fact that in different regions of hidden layer state space a small change in activation may have more or less effect on the output activation (a modification I call processing topography analysis).

The proposal also gives rise to an empirical prediction (for those readers who think

---

[28]    Rolls & Treves (1998).

[29]    Only simple linear combinations are envisaged, e.g: a, b, and a + b; with all or most of the principal components corresponding to clusters that are actually realised in the state space – not just any reparameterisation will do.

that entails that this can't be philosophy: look away now). The microfeatural idea motivates another standard way to investigate the internal workings of a trained network: knock out one of the hidden layer nodes, and from the pattern of error which results from the lesion, infer the representational role of that node. Since clusters are independent of hidden layer nodes, no analogue of a physical lesion is available to knock out a single cluster. However, something similar is accomplished by removing the effect of a single hidden layer cluster. This is achieved by subtracting from the hidden layer activation vector the component (if any) parallel to that cluster, before activation is passed on to the output layer. (That effect can be implemented across the board by a suitable transformation of the matrix of weights between hidden layer and output layer.) My prediction is that after such a notional lesion the network will tend to mis-classify samples which have the property that was represented by the cluster that has been artificially removed.

## 8. Conclusion

In the debate between philosophers about whether states of connectionist systems are contentful the tacit assumption that the vehicles of content are particular distributed patterns of activation has been left unexamined. If the vehicles of content in connectionist systems are clusters in state space, then it is much easier to see that some connectionist networks operate by processing internal representations.

*Nicholas Shea*

*Faculty of Philosophy*

*University of Oxford*

Word count: 10,815

**References**

Bechtel, W. and A. Abrahamsen 2002: *Connectionism and the Mind*, 2nd ed. Oxford: Blackwell.

Calvo Garzón, F. 2003: Connectionist Semantics and the Collateral Information Challenge. *Mind & Language,* 18(1), 77-94.

Churchland, P. M. 1981: Eliminative Materialism and the Propositional Attitudes. *Journal of Philosophy,* 78, 67-90.

Churchland, P. M. 1998: Conceptual Similarity Across Sensory and Neural Diversity: The Fodor/Lepore Challenge Answered. *Journal of Philosophy,* 95(1), 5-32.

Churchland, P. M. (in progress): *Inner Spaces and Outer Spaces: The New Epistemology.*

Churchland, P. S. and T. J. Sejnowski 1989: Neural Representation and Neural Computation. In Nadel, Cooper, Culicover & Harnish (eds.), *Neural Connections, Mental Computations.* Cambridge, Mass: MIT Press.

Churchland, P. S. and T. J. Sejnowski 1992: *The Computational Brain.* Cambridge, Mass: MIT Press.

Clark, A. 1993: *Associative Engines.* Cambridge, Mass: MIT Press.

Clark, A. 1996: Dealing in Futures: Folk Psychology and the Role of Representations in Cognitive Science. In R. McCauley (ed.), *The Churchlands and Their Critics.* Cambridge, Mass: Blackwell.

Clark, A. 2001: *Mindware*.  Oxford: O.U.P.

Cottrell, G. W. and J. Metcalfe 1991: Empath: Face, gender and emotion recognition using holons. In Lippman, R. P., Moody, J., and Touretzky, D. S., (eds), *Advances in Neural Information Processing Systems 3*, 564–571. San Mateo, Morgan Kaufmann.

Cummins, R. 1997: The lot of the causal theory of mental content. *Journal of Philosophy,* 94(10), 535-542.

Dawson, M. and C. D. Piercey 2001: On the Subsymbolic Nature of a PDP Architecture that Uses a Nonmontonic Activation Function. *Minds and Machines,* 11, 197-218.

Dawson, M. et al 2000: Using extra output learning to insert a symbolic theory into a connectionist network. *Minds and Machines,* 10, 171-201.

Dennett, D. C. 1987: *The Intentional Stance*. Cambridge, MA, MIT Press.

Eliasmith, C. 2003: Moving beyond metaphors: understanding the mind for what it is. *Journal of Philosophy*, 493 – 520.

Elman, J. 1990: Finding structure in time. *Cognitive Science* 14, 179-212.

Elman, J. 1991: Incremental Learning, or the Importance of Starting Small.  Technical report 9101, Center for Research in Language, U.C.S.D.  Described in detail in Clark 1993, pp. 138-142.

Fodor, J. & E. Lepore 1999: All at sea in semantic space: Churchland on meaning similarity. *Journal of Philosophy,* 96(8), 381-403.

Fodor, J. 1998: *Concepts: Where Cognitive Science Went Wrong*. Oxford: Clarendon Press.

Fodor, J. 2000: *The Mind Doesn't Work That Way* Cambridge, MA, MIT Press.

Fodor, J. A. and Z. W. Pylyshyn 1988: Connectionism and Cognitive Architecture: A Critical Analysis. *Cognition* 28, 3-71.

Hinton, G. 1989: Connectionist learning procedures. *Artificial Intelligence* 40, 185-234.

Karmiloff-Smith, A. 1994: Précis of: *Beyond Modularity: A developmental perspective*. *Behavioural and Brain Sciences,* 17(4), 693-745.

Laakso, A. & G. Cottrell 2000: Content and cluster analysis: assessing representational similarity in neural systems. *Philosophical Psychology,* 13(1), 47-76.

Laurence, S. and E. Margolis 2002: Radical Concept Nativism. *Cognition,* 86, 25-55.

Lehky, S. and T. Sejnowski 1987: Extracting 3-D curvatures from images using a neural model, *Society for Neuroscience Abstracts,* 13, 1451.

Lehky, S. and T. Sejnowski 1988: Neural network model for the representation of surface curvature from images of shaded surfaces. In J. Lund (ed.), *Organising Principles of Sensory Processing*. Oxford: O.U.P.

Mozer, M. and P. Smolensky 1989: Using relevance to reduce network size automatically. *Connection Science,* 1(1), 3-17.

O'Brien, G. & Opie, J. 2001. Connectionist Vehicles, Structural Resemblance, and the Phenomenal Mind.  In J. Veldeman (ed.), Naturalism and the Phenomenal Mind, a special issue of *Communication and Cognition,* 34, 13-38.

Pollack, J. 1990: Recursive distributed representations. *Artificial Intelligence,* 46, 77-105.

Ramsey, W., S. Stich and J. Garon 1990: Connectionism, Eliminativism, and the Future of Folk Psychology. *Philosophical Perspectives* 4, 499-533.  Repr. in *Connectionism: Debates in Psychological Explanation*, C. MacDonald and G. MacDonald (eds.), 1995 Oxford: Blackwell.

Rolls, E. and A. Treves 1998: *Neural Networks and Brain Function*. Oxford, OUP.

Rupert, R. 2001: Coining Terms in the Language of Thought. *Journal of Philosophy,* 499-530.

Sejnowski, T. and C. Rosenberg 1987: Parallel Networks that Learn to Pronounce English Text. *Complex Systems,* 1, pp. 145-168.

Smolensky, P. 1988: On the Proper Treatment of Connectionism. *Behavioural and Brain Sciences,* 11, 1-74.

Tiffany, E. 1999: Semantics San Diego Style. *Journal of Philosophy,* 96, 416-429.

Usher, M. 2001. A Statistical Referential Theory of Content: Using Information Theory to Account for Misrepresentation. *Mind & Language,* 16(3), 311-334.