Running Head: Hyperspecificity, Autism, and Neural Nets

The Basis of Hyperspecificity in Autism:

A Preliminary Suggestion Based on Properties of Neural Nets

James L. McClelland

Carnegie Mellon University and the

Center for the Neural Basis of Cognition

Send Correspondence to:

James L. McClelland

Department of Psychology

Carnegie Mellon University

Pittsburgh, PA 15213

jlm@andrew.cmu.edu

(412)-268-3157 (Voice) / (412)-268-5060 (Fax)

Abstract

This article reviews a few key ideas about the representation of information of information in neural networks, and uses these ideas to address one aspect of autism, namely the apparent hyperspecificity that is often seen in autistic children's application of previously acquired information. Hyperspecificity is seen as reflecting a possible feature of the neural codes used to represent concepts in the autistic brain.

This short paper presents some initial thoughts on autism stemming from my own particular background and perspective. I come from a computational tradition called the connectionist, parallel distributed processing or neural network tradition, wherein we view cognitive processes as arising from the interactions of large numbers of simple processing units (Rumelhart, McClelland, & the PDP Research Group, 1986b). In this approach, the representation of an item when active is a pattern of neural activity over the units in a network, and processing occurs through the propagation of activity among the units via weighted connections. Learning occurs through changes to the connection weights, and results in changes in the representation of information and the processing that takes place in response to particular inputs.

In previous work, I and others have applied these ideas to a number of phenomena in normal adult cognition, disorders of cognition, and in cognitive development. The present paper represents some initial thoughts about applying this framework in an attempt to understand certain aspects of autism. I am far from an expert in on autism, and thus these thoughts must be considered highly preliminary. However, it is my impression that some of the ideas we have developed can be helpful in understanding some of the features of autism, as these are reported in review articles such as those by Rumsey (1992) and Happe and Frith (1996). Others (Cohen, 1994; Gustafsson, 1997) have also offered insights into aspects of autism based on related neural network approaches. There is considerable common ground between my ideas and the suggestions of these other authors.

In what follows, I will consider some key aspects of representation and learning in neural networks and apply these to a consideration of one specific aspect of autism, namely the apparent hyper-specificity of autistic children's memories. This aspect is well brought out by the poignient example of a young autistic child who was at a birthday party and had to go to the potty. Apparently he didn't recognize the potty at the home of the child who was having the party, since

it was different from the one at home, and thus he refused to use it. The only resolution was to leave the party and take the child home.

Such examples suggest that in the mind of the autistic child, there is a tendency to represent things in an extremely specific way. This this might have some relationship to the literal-mindedness of autistic children, their lack of sensitivity to meaningfulness in memory tasks, but also – and we will see how this comes out of the network models – their relatively spared abilities in certain kinds of memory, particularly very rote forms of memory where meaningfulness and generalization is less of an issue. My main goal is to give you a sense of how a neural network theorist might have a perspective on that aspect of autism.

Figure 1 shows a neural network; a very generic one. It has two sets of units and connections from one to the other set. Let us think about learning in this network. As already noted, this occurs through changing the strengths of the connections between the neurons so as to allow one pattern of activation—on the input units—to evoke another—on the output units. What I want to illustrate is how the storage of memories in a neural network leads simultaneously to two related characterisics: generalization and interference. These characteristics lie at the heart of the problem that the brain must solve if it is going to be effective in storing information in memory.

Insert Figure 1 about here.

As a concrete example, you might think of the sight of a rose being paired with the aroma of the rose. In this case, what we want to do is set the weights, so that the sight of the rose (on the evokes a kind of a memory of what that rose smelled like. If we imagine the sight of the rose activates the neurons in the vertical row on the left and the smell activates the neurons in the horizontal row at the bottom.

We can create an association between the sight and the aroma if we strengthen the connections from the active input neurons to the active output neurons. This was Hebb’s formulation of how

learning occurs in the brain: He suggested that if neuron A participated in the firing of neuron B, the strength of the connection between them was increased. The effect of this is to allow the sight of the rose to elicit a memory of its aroma.

One thing that we can observe is that this kind of memory is capable of storing several associations. For example, I could show a beef steak to somebody and let them sniff it and thereby produce different visual and aroma representations. We could also set the connection weights to capture that association. However, as soon as you start to think about what would happen if I tried to learn multiple associations, you begin to notice that there is potentially some cross-talk, i.e, some interaction between the memories. These interactions are the essence of how neural networks operate.

In particular, interesting things begin to happen whenever two input patterns overlap. In this case, what happens is that when I present either pattern, it will tend to engage some of the knowledge that I acquired in learning the other one. Thus, if the steak input and the rose input overlap at all, there will be a tendency for the output to include some of both their associations. This is illustrated in the figure for two arbitrary input-output pattern pairs.

Pattern overlap creates a basis for the shared use of knowledge that was acquired in the neural network. Now, that canbe very useful. In particular, if two inputs are very similar, the neural network will tend to react to them in similar ways. If there are several different overlapping examples, and if each time we experience any one of these examples, we make slight increases in the connection weights between active units, then the common elements will tend to be strengthened repeatedly. In this way, we can gradually build up a representation of the association that produces a robust response whenever any input that contains most of the common elements is present.In other words, the network will be able to generalize to other inputs sharing similar elements.

Although this property of generalization is very useful, but it can also lead to problems. For example, it could be a serious problem for associating names with faces. After all, whereas faces are often rather similar to each other, everybody has a different, unique name. Problems are created if every time we see a face that is similar to another one that we have seen many times, we tend to give it that other person's name. So, potential benefits in the form of generalization are countered by costs in the form of interference from the tendency of neural networks to share what they've learned across patterns.

Now, there are ways of reducing the degree to which patterns of activation that are used as the input to a learning process overlap. A very common idea is to recode the input patterns, using a technique called conjunctive coding, an idea illustrated in Figure 2. We'll start with a representation that shows overlap between two similar inputs, then solve the problem using conjunctive coding. Let's imagine that we are representing objects with simple patterns of activity over a population of neurons, as illustrated on the lower part of the figure. I have made the patterns very simple ones, with just two neurons active in each. We might imagine the neurons represent distinct features of a single object. For example, if the object were a red square, it migth activate one unit representing red and anoth other representing square. Similarly, for a red circle, it might activate the same red unit and another unit representing circle. Using this approach, we would get some overlap in the representations, since in both cases the unit for red will be active. But we can defeat that tendency completely if we use conjunctive coding, in which we assign neurons to become active only when the particular combination of elements occurs. We can arrange this by having every pair of neurons in the lower layer shown in the figure project to an unique neuron that represents the simultaneous occurrence both features together. To do this, we need connections from each of the two inputs, and we need to arrange things so that the neuron does not fire unless both inputs are present. One can arrange this by assigning a threshold to each

of the neurons in the upper layer, so that it will not fire unless both of its inputs are active. There are other ways, but I will not go into those details here.

Insert Figure 2 about here.

The key point is that if we have a neuron that represents the conjunction of red and square, then we could learn associations to the activity of this neuron that would be specific to that conjunction. When a red circle is presented, the unit will not be activated, and so any response based on outgoing connections from this neuron will not be activated. Thus, through the use of conjunctive coding, we can minimize the extent to which our knowledge in the neural network can be applied to overlapping inputs. Note that this scheme can be made considerably more sophisticated, and there is a vast literature on various neural network schemes for conjunctive coding. One general version of this idea, that of David Marr (1969), suggests that each conjunctive unit (called a Codon) receives inputs from some particular number R (in the brain the number could be say 5000) input units, and reponds when some smaller number S (say 250) of the inputs are active. Such a scheme can be used to recode patterns with many active elements (possibly a few thousand out of a million). Depending on the values of R and S, this can drastically reduce the overlap of incoming patterns of activation. Another, similar scheme is described in O'Reilly and McClelland (1994).

We can now think about the bathroom that that little autistic child had in his house as being in a room in which the walls might have been green and the tile on the floor might have been black and white and the shower curtain was perhaps a map of the world. And he goes into a very different bathroom at the birthday party with none of these features. There is still a potty and a sink, perhaps, but none of those other properties are there. Perhaps what differs between this autistic child and the rest of us is that the autistic child associates using the potty with a highly conjunctive representation, whereas the rest of us use representations that tend to preserve the

individual elements—we use what we often call a componential representation. If this is so, it would go some way toward accounting for the difference between children with autism and other healthy children.

With this background, let me take this a step further and talk about an issue that arises when neural network researchers think further about cognitive development in the brain. We think about cognitive development as being a process that involves the discovery of ways to represent experiences we have with the world in a way that is helpful. We learn to represent information in order to exploit generalizations when appropriate, and to avoid generalization when inappropriate. It is likely that brain systems differ in their tendency to use componential or conjunctive representations, and this is helpful; but in addition, we rely on learning to help establish representations that are useful in different domains.

To me, this tendency of networks to discover useful representations is the most interesting aspect of the discovery in the 1980's of powerful methods for adjusting the connections in neural networks (e.g. Rumelhart, Hinton, & Williams, 1986a; Ackley, Hinton, & Sejnowski, 1985). These methods allow networks to gradually discover useful representations that help pull out particular conjunctions that are useful (the potty itself, as a complex object, is certainly a conjunction of elements), collapse together representations of things that do not need to be distinguished, and assign highly distinct representations to other things that may be superficially similar but that must be told apart.

For example, Rumelhart and Todd (1993) and later McClelland, McNaughton, and O'Reilly (1995) studied a neural network model that they trained to answer questions about various different kinds of things, as illustrated in Figure 3. In this network, there are units for each of several different kinds of things, some birds, some fish, some flowers, and some trees. There are also units for different kinds of queries, labeled IS A, CAN, IS, and HAS. The task of the net is to activate appropriate answer units when given a probe, such as 'ROBIN CAN'. In this case, it

should activate 'GROW MOVE FLY' since these are the things that robins can do. The network is trained by presenting many probes, letting it generate an output, then adjusting the weights in the network to reduce the difference between the obtained and the desired output. On each learning trial, only tiny adjustments are made, so that the network learns very slowly. As it progresses, it adjust the weights in all parts of the network. The weights from the input to what I have called the concept representation units are the most important ones for our purposes; they allow the network to assign useful representations to each of the concepts.

Insert Figure 3 about here.

What happens as learning occurs is that the representations in the network are gradually differentiated, pulling apart those concepts that are very different but allowing those that are very similar to have highly overlapping representations (See Figure 4. This is useful, because once the representations are established, what we learn about one kind—say the robin—tends to automatically generalize to other similar kinds, say the bird. But, because the representation Robin is very different than the representations of all of the plants, what we learn about robins does not generalize to any of the plants.

Insert Figure 4 about here.

Although very simple, this network illustrates what in my view is the key process that occurs during cognitive development: the discovery of appropriate representations that allow effective generalization. This process allows networks to capture the tendency seen in humans to apply progressively finer and finer distinctions to conceptual knowledge as they grow older Keil (1979).

With these aspects of neural networks in mind, we can now complete the thought about hyper-specificity in autism. The suggestion is simple: due to what may perhaps be a very subtle

change in some of the parameters of the real neural networks in the brains of children with autism, they may be predisposed to use an excessively conjunctive form of neural coding, at least in certain parts of their brains (particularly those devoted to semantic and conceptual representations). This could prevent them from exploiting overlap in cases where overlap leads to the useful ability to generalize. Instead, it would leave the child with the ability to learn associations to particular, specific inputs and without the ability to extend what they have learned to other similar experiences.

References

Ackley, D. H., Hinton, G. E., & Sejnowski, T. J. (1985). A learning algorithm for Boltzmann machines. *Cognitive Science, 9*, 147–169.

Cohen, I. L. (1994). An artificial neural network analogue of learning autism. *Society of Biological Psychiatry*, 5–20.

Gustafsson, L. (1997). Inadequate cortical feature maps: A neural circuit theory of autism. *Society of Biological Psychiatry*, 1138–1147.

Happe, F., & Frith, U. (1996). The neuropsychology of autism. *Brain, 119*, 1377–1400.

Keil, F. C. (1979). *Semantic and conceptual development: An ontological perspective*. Cambridge, MA: Harvard University Press.

Marr, D. (1969). A theory of cerebellar cortex. *Journal of Physiology (London), 202*, 437–470.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). 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*, 419–457.

O'Reilly, R. C., & McClelland, J. L. (1994). Hippocampal conjunctive encoding, storage, and recall: Avoiding a tradeoff. *Hippocampus, 4*, 661–682.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986a). Learning internal representations by error propagation. In D. E. Rumelhart, J. L. McClelland, & the PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition*, Vol. 1 (pp. 318–362). Cambridge, MA: MIT Press.

Rumelhart, D. E., McClelland, J. L., & the PDP Research Group (1986b). *Parallel distributed processing: Explorations in the microstructure of cognition. Volume I: Foundations & Volume II: Psychological and biological models*. Cambridge, MA: MIT Press.

Rumelhart, D. E., & Todd, P. M. (1993). Learning and connectionist representations. In D. E. Meyer, & S. Kornblum (Eds.), *Attention and performance XIV: Synergies in experimental*

psychology, artificial intelligence, and cognitive neuroscience (pp. 3–30). Cambridge, MA: MIT Press.

Rumsey, J. (1992). High-functioning individuals with autism. In E. Schopler, & G. B. Mesibov (Eds.), *Neuropsychological studies of high-level autism* (pp. 41–64). New York: Plenum Press.

Figure Captions

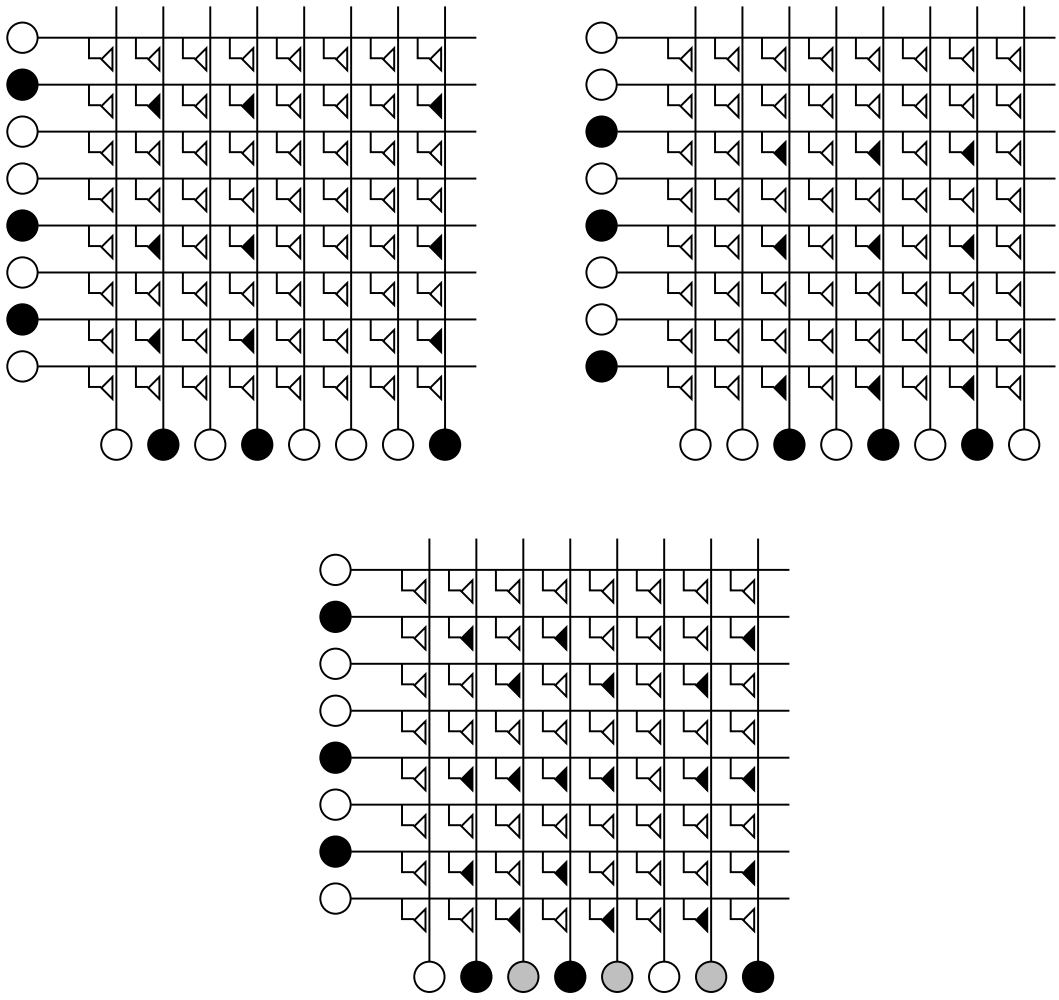
Figure 1. Effects of learning two pairs of input-output patterns (above) on the response to one of the two patterns. Since both associations are stored in the connection weights, and since the input patterns overlap, the outputs tend to contain some of both of the patterns stored in the connection weights. This sort of cross-talk between patterns can be minimized by reducing input pattern overlap.

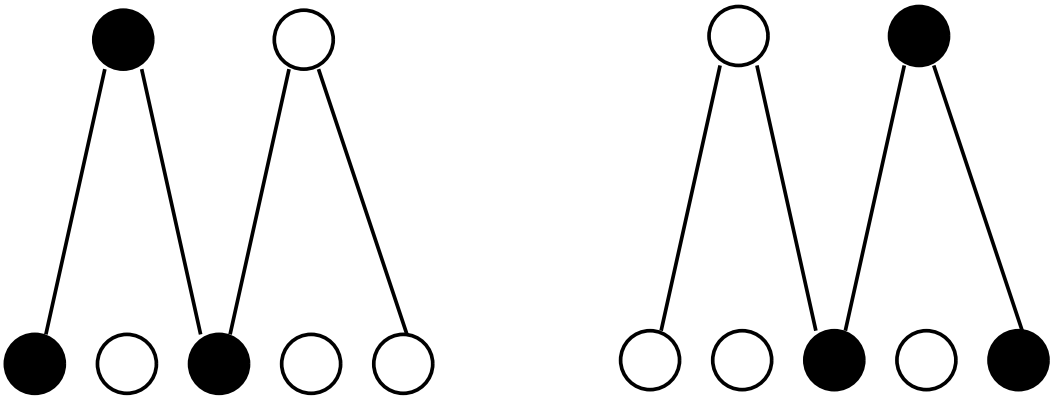
Figure 2. Using conjunctive coding to reduce pattern overlap. The same small network of five input units and two conjunction units is shown twice. The input pattern on the left and the input pattern on the right are partially overlapping—they share one unit in common. By using a conjunctive code, in which individual neurons are dedicated to particular pairs of inputs, the overlap can be eliminated (each input pattern only activates one conjunction unit).

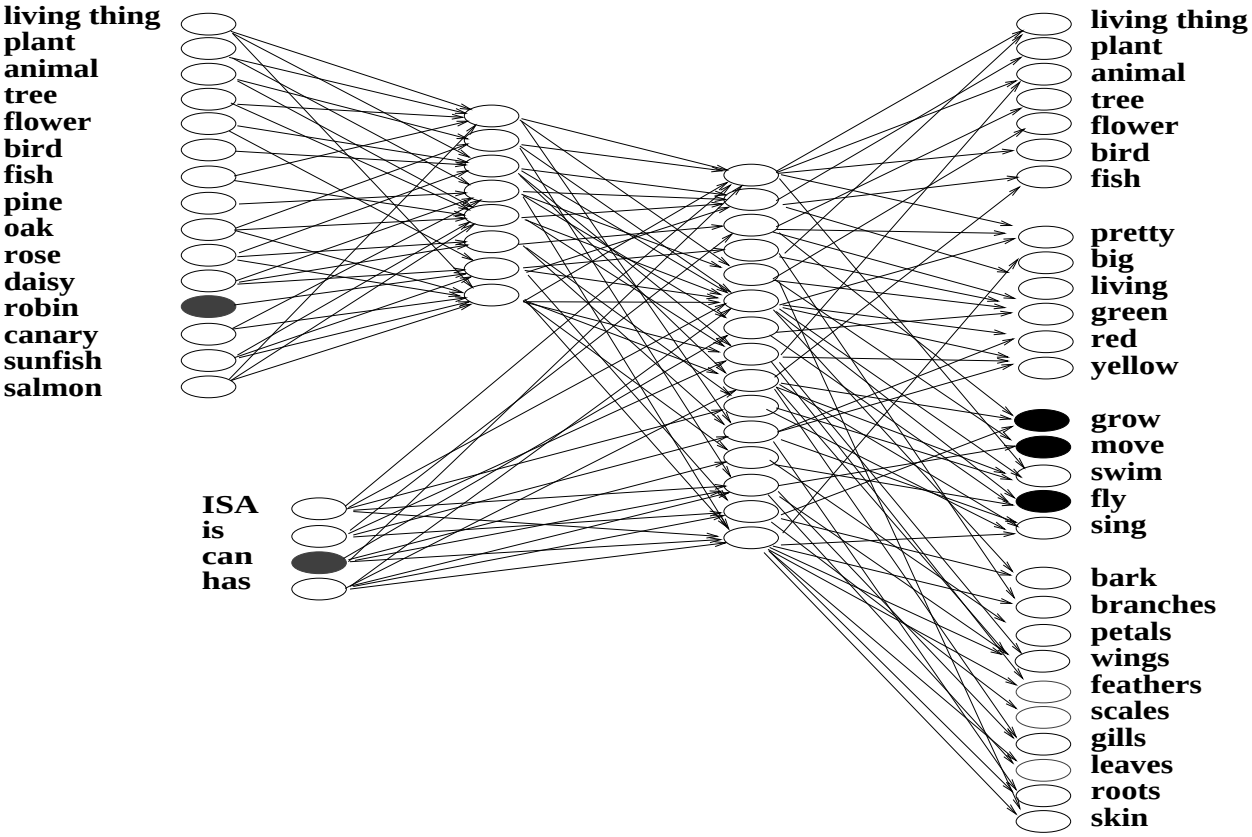
Figure 3. Our depiction of the connectionist network used by Rumelhart to learn propositions about several different kinds of plants and animals. The entire set of units used in the actual network is shown. Inputs are presented on the left, and activation propagates from left to right. Where connections are indicated, every unit in the pool on the left (sending) side projects to every unit in the right (receiving) side. An input consists of a concept-relation pair; the input *robin can* is illustrated here by darkening the active input units. The network is trained to turn on all those output units that represent correct completions of the input pattern. In this case, the correct units to activate are *grow, move* and *fly*; the units for these outputs are darkened as well. Subsequent analysis focuses on the concept representation units, the group of eight units to the right of the concept input units. *Note* Figure is based on the network depicted in Rumelhart and Todd (1993), Figure 1.9, page 15. Reprinted with permission from Figure 5 of “Why there are complementary learning systems in hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory” by J. L. McClelland, B. L. McNaughton, and R.

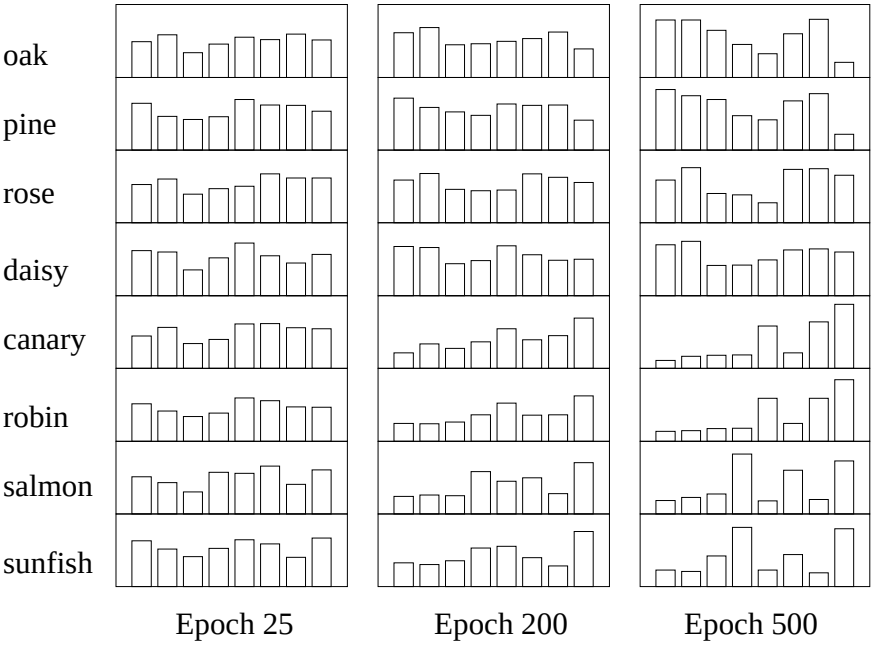
C. O'Reilly, 1995, *Psychological Review*, p. 430. Copyright 1995 by the American Psychological Association.

Figure 4. Representations discovered in training the network on facts about the plants and animals, using the network shown in Figure 3. The figure presents a vertical bar indicating the activation of each of the eight concept representation units produced by activating the input unit for each of the eight specific concepts. The height of each vertical bar indicates the activation of the corresponding unit on a scale from 0 to 1. One can see that initially all the concepts have fairly similar representations. After 200 epochs, there is a clear differentiation of the representations of the plants and animals, but the trees and flowers are still quite similar as are the birds and the fish. After 500 epochs, the further differentiation of the plants into trees and flowers and of the animals into fish and birds is apparent. Reprinted with permission from Figure 6 of "Why there are complementary learning systems in hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory" by J. L. McClelland, B. L. McNaughton, and R. C. O'Reilly, 1995, *Psychological Review*, p. 431. Copyright 1995 by the American Psychological Association.






living thing
plant
animal
tree
flower
bird
fish
pine
oak
rose
daisy
robin
canary
sunfish
salmon
ISA
is
can
has
living thing
plant
animal
tree
flower
bird
fish
pretty
big
living
green
red
yellow
grow
move
swim
fly
sing
bark
branches
petals
wings
feathers
scales
gills
leaves
roots
skin


oak
pine
rose
daisy
canary
robin
salmon
sunfish
Epoch 25
Epoch 200
Epoch 500