Brain and Mind **2:** 161–194, 2001.
© 2001 *Kluwer Academic Publishers. Printed in the Netherlands.*

The New Associationism: A Neural Explanation for the Predictive Powers of Cerebral Cortex

DAN RYDER[1]* and OLEG V. FAVOROV[2]
[1]*Department of Philosophy, C.B. #3125, University of North Carolina at Chapel Hill, NC 27599-3125, U.S.A. (E-mail: dan@danryder.com);* [2]*School of Electrical Engineering and Computer Science, University of Central Florida, Orlando, FL 32816, U.S.A. (*author for correspondence)*

(Received: 19 July 2000; in final form: 23 May 2001)

Abstract. The ability to predict is the most important ability of the brain. Somehow, the cortex is able to extract regularities from the environment and use those regularities as a basis for prediction. This is a most remarkable skill, considering that behaviourally significant environmental regularities are not easy to discern: they operate not only between pairs of simple environmental conditions, as traditional associationism has assumed, but among complex functions of conditions that are orders of complexity removed from raw sensory inputs. We propose that the brain's basic mechanism for discovering such complex regularities is implemented in the dendritic trees of individual pyramidal cells in the cerebral cortex. Pyramidal cells have 5–8 principal dendrites, each of which is capable of learning *nonlinear* input-to-output transfer functions. We propose that each dendrite is trained, in learning its transfer function, by all the other principal dendrites of the same cell. These dendrites teach each other to respond to their separate inputs with *matching* outputs. Exposed to different but related information about the sensory environment, principal dendrites of the same cell tune to *functions* over environmental conditions that, while different, are *correlated.* As a result, the cell as a whole tunes to the source of the regularities discovered by the cooperating dendrites, creating a new representation. When organized into feed-forward/feedback layers, pyramidal cells can build their discoveries on the discoveries of other cells, gradually uncovering nature's hidden order. The resulting associative network is powerful enough to meet a troubling traditional objection to associationism: that it is too simple an architecture to implement rational processes.

Key words: associative learning, cerebral cortex, dendritic function, mental representation, plasticity, reasoning

1. Introduction: The Brain's Task

In 1988, neural network modeler Jerome Lettvin had this to say about the state of the sciences of the mind (p. v):

> The literature in neuroscience and psychology is now so large and growing so fast that no one can master what has been written or keep up with current work. That is because it is composed of endless empirics. There is no theory to give coherence to this mountain of data in terms of how the brain functions as a mechanism that sustains mental process. Since, in the end, that is the goal of

> neuroscience – to account for how such process can occur – the absence of working models points out how undeveloped is the discipline.

Now that the decade of the brain is over, and we look back at the progress we have made since Lettvin claimed neuroscience was in its infancy, can we at least say that we have reached puberty? Has the discipline gone past endless empirics, and achieved some real understanding? Most will agree that such advances have been, at best, few and far between. Neuroscientific theories that can also be considered theories of the *mind* are exceedingly rare.

There may be a good reason for this. Perhaps the brain is a collection of specialized mechanisms that have no interesting functional properties in common, and the mind somehow 'emerges' from this hodgepodge, as Rodney Brooks (1991), Daniel Dennett (1994), and Steven Pinker (1997) would have us believe. Then it would be no wonder that neuroscience has not produced a general theory of the mind/brain. We take a more hopeful view. We, among others (Braitenberg, 1978; Mountcastle, 1978; Edelman, 1987; Burnod 1988; Abeles, 1991; Barlow, 1992; Phillips and Singer, 1997; Quartz and Sejnowski, 1997; Grossberg, 2000) believe that there is a single, unifying task that the cerebral cortex performs, and that an understanding of the nature of the task and how the cortex manages to perform it is an important step towards understanding how the mind and the brain can be one.

What is this task? Our bet is that it concerns *prediction*. The brain must choose actions on the basis of *present* environmental conditions, but for their *future* effects. The better an organism is able to predict the future course of events, the better will it be able to adjust its behaviour to respond to and influence its environment. This much is clear. What is unclear is how this predictive ability is implemented in humans and our relatives. As this question is ultimately an empirical one, in this paper we propose a testable neural network model, grounded in neurobiology, for how the cerebral cortex accomplishes the all-important task of prediction.

Some of an organism's predictive abilities are endowed by neural mechanisms that are instinctive, acquired by the species through the process of natural selection. More flexibly, and more importantly for advanced animals, other predictive abilities are acquired by the individual through the process of learning. We adhere to the view that this acquisition is a process of associative learning, where regularities in the environment – as reflected in the patterns of an organism's receptor activities – are extracted by an associative device.

The classical empiricist account of how the mind acquires and uses its predictive powers is associationist. According to this account, if events A and B regularly occur simultaneously or successively, representations of these events in the mind of an observer become 'associated,' such that the thought of A calls to mind the thought of B. In future, upon observing A, this 'calling to mind' makes the agent expect B; this is the classical empiricist's explanation of prediction. That is, association occurs as a result of correlation, and correlation may be used for prediction in the absence of one of the correlates. David Hume (1740/1978) was the first to propose this as a fully general account of the mind's operation. According to

Hume, his proposed laws of association were the basic laws of psychology, just as Newton's laws of motion were the basic laws of physics. Hume's associationism was thus a *mechanistic* theory of the mind. By this, we do not mean that Hume was proposing a physical explanation of the mind, for he was not. His account was, however, a reductive and broadly naturalistic one. Prior to Hume, *reason* was taken for granted as a basic, unanalyzable essence of mind. Associationism was meant to reduce reason to the non-rational. It is in this sense that Hume's theory was mechanistic. (David Hartley (1749/1970) later proposed a physical implementation for Hume's mechanism.)

The fundamental problem with Hume's theory is that pairwise association is not a sufficiently powerful basic element with which to construct the operations of reason (Fodor, 1983, pp. 23–36; Fodor and Pylyshyn, 1988; Rosenberg, 1997; Rey, 1997). A criticism originally leveled at Hume by Kant (1787/1996) is repeated for the benefit of modern associationists by Jerry Fodor: "The traditional, fundamental, and decisive objection to association is that it is too stupid a relation to form the basis of a mental life." While modern associative artificial neural networks may be much smarter than anything Hume ever thought of, the Kant/Fodor objection is still depressingly apt (e.g., Minsky and Papert, 1988; Clark and Thornton, 1997). Could this explain why neuroscientists have failed to produce models of the mind, as opposed to models of the brain? It is not Hume's mechanistic theory, but Turing's, that has had some success in modeling reason. But the brain rarely plays more than a small part (if any) in such theories, and they do not sit well with empiricism. In fact they form the basis of the hodgepodge view we hope to avoid, namely multifarious modules each performing their own proprietary algorithm.

Another problem with Hume's theory, if one's goal is naturalizing the mind, lies with the second basic operation (besides association) that Hume postulates. This is *abstraction*, whereby the mind creates 'ideas', representations, or concepts of features in its environment that are not reflected directly in sensation (1740/1978, I,I,vii). Unless the associationist believes we are innately stocked with an exhaustive set of representations, some such mechanism of representation acquisition will be required. The problem with Hume's operation of abstraction, from the naturalist's point of view, is that it is entirely unexplicated. While his account of association is mechanistic, his account of abstraction is not. All he says is that we abstract out those features of objects and events that are similar from one occasion to the next.[1] He thinks that wherever there is similarity, it causes us to abstract a concept of the feature in common. This must be false, however,

[1] Hume's account of abstraction is a bit more complicated than this suggests. It goes like this: one notices a similarity among several ideas of particulars, and the ideas of these particulars become "annexed" to a general term. In reasoning, one makes use only of ideas of specific particulars, but they are made to signify many other particulars by virtue of their connection to the general term. It is really the rule (or "habit" as Hume calls it) that links general terms with sets of particular ideas that serves the function of an abstract concept. (For instance, one particular idea can be linked to two different general terms by distinct 'habits.') Stroud shows how the role played by these 'habits' in reasoning violates naturalism (1977, pp. 37–41).

since there are an indefinite number of similarities amongst any random sample of objects, and we only abstract out a vanishingly small percentage of these. Which ones? It seems that the mind chooses to represent the *useful* ones, those features that are predictively related to other features; this is rational, if these representations are then to figure in associations. But Hume gives us no mechanistic account of the apparently rational operation of abstraction itself. Hume virtually acknowledges the problem when he says that the collection of ideas occurs by some "magical faculty in the soul," which is highly developed in geniuses.

In this paper, we propose a new associationism, an updated version of Hume's empiricist theory. We present a 'smart' though fully naturalistic associationism. The basic mechanism we propose, implemented (we believe) at the level of single cells in the cerebral cortex, is sufficiently powerful to form the building blocks of reason. At the same time, it replaces Hume's operation of abstraction with a mechanism. In fact, the mechanism of association and the mechanism that replaces abstraction turn out to be identical, which results in a unified explanation of two fundamental mental processes: rational transitions in thought (reasoning) and representation acquisition. This yields the beginnings of a neural theory, not only of the brain, but also of the mind.[2] We start by considering one fundamental way in which Hume's associationism needs to be enriched.

2. Discovering Correlated Functions over Different Sets of Conditions

Hume believed that grasping nature's order was a matter of finding correlations between pairs of individual conditions. But the regularities among natural phenomena that are most useful behaviourally do not typically operate just between pairs of individual conditions. Rather, they operate among combinations of multiple conditions, such that what are correlated are *functions* over non-overlapping subsets of these conditions. Such functions might be logical, or mathematical (nonlinear in particular), or of any other kind. Once one begins to see the complexity of the correlations that would need to be reflected in a mechanism of association, it is clear that they are ubiquitous. The stock-in-trade example of a simple correlation is thunder with lightning. But what ought one to associate with, say, the conjunctive condition of a small movement in a curtain on your bedroom window? In the condition where the window is open, it could be the wind, or a cat (whether or not one owns a cat), or a person, or a heating vent (in the winter). In the condition where the window is closed, it could not be the wind or a person; however it could be a cat (but only if one owns a cat), or a heating vent (in the winter). You will notice that even this example is over-simplified, and also that it is very easy to generate such examples. The simple thunder-lightning association is not the rule. Thus it is clear that Hume's associationism needs to be enriched so

[2] A third fundamental process needs explanation, namely decision making and action. This is one focus of our current research.

that not only pairs of individual conditions can be associated, but functions over multiple conditions as well.

Learning to appreciate these predictive relations among functions of conditions is an extraordinarily difficult task, as there are an infinite number of possible functions that can be defined over sets of conditions, most of which will not yield useful regularities. This is the problem with Hume's abstraction that he failed to appreciate, but it is exacerbated exponentially when the field of candidate associable conditions includes functions. Unless we opt for a nativist theory that says we are born being able to recognize just the right functions over sets of conditions, this problem can only be solved by a 'smart' associative device.

The computational approach to finding correlated functions over different inputs was originally formulated by Becker and Hinton (1992) and developed further by Becker in a series of papers (1995, 1996, 1999). The basic idea is to set up two or more computational modules that look at separate but related parts of the sensory input, and to make these modules teach each other to produce outputs that match as closely as possible. In this way the modules will learn to perform different operations on their (differing) inputs so as to yield consistently correlated outputs.

Becker and Hinton's basic concept forms one of the cornerstones of Phillips and Singer's (1997) view of the cerebral cortex. Phillips and Singer propose that contextual (i.e., lateral, as oppose to afferent) inputs to cortical cells guide them to tune to those stimulus features in their receptive fields that are predictably related to the context in which they occur. In this way, by discovering predictable relations between different inputs (within vs. outside their receptive fields), cortical cells discover objectively important variables in the external world.

By proposing that cortical cells tune preferentially to those features of their sensory environments that are predictably related to other such features, Phillips and Singer (1997) advanced a new important guiding principle for cortical cells' functional tuning. They also offered a possible mechanism for how such tuning might be accomplished. Unfortunately, this mechanism is severely limited in its ability to discover predictable relations between *nonlinear* functions of environmental conditions. This weakness of their mechanism prompted Phillips and Singer to conclude that the "reliable discovery of such nonlinear variables may not be a fundamental capability of cortex" (p. 709). They noted that error-backpropagation learning would be more powerful, but did not consider its use in the cortex to be a real possibility.

Phillips and Singer's focus on individual pyramidal cells as discoverers of correlated functions is very appealing for a number of reasons. First, placing the mechanism in the dendritic trees of individual cells (rather than requiring multi-cellular modules) expands by orders of magnitude the number of regularities that the brain will be able to discover in the environment and makes this mechanism much more practical biologically. Second, the neocortex is the part of the brain primarily engaged in the task of discovering regularities (Barlow, 1992), and pyramidal cells are the main type of neurons that compose it. Pyramidal cells have 5–8 principal dendrites growing from the soma, including 4–7 basal dendrites, and the apical dendrite with its side branches (Feldman, 1984). Each principal dendrite sprouts an elaborate, tree-like pattern of branches and is capable of complex forms of integration of synaptic inputs it receives from other neurons (Mel, 1994). Through its synapses, each principal dendrite receives information from other neurons about different environmental conditions. Exposed to this information during different situations experienced by the animal, each principal dendrite gradually learns (by adjusting its synaptic connections) to respond in a particular way to patterns of synaptic inputs it receives (Singer, 1995).

The third consideration is that synaptic learning in dendrites is controlled both by presynaptic activity and by the output activity of the postsynaptic cell; the latter is apparently signaled to each synapse by spikes that are backpropagated from the soma up through each dendrite (Markram *et al.*, 1997b; Stuart *et al.*, 1997). The output activity of a cell itself is a sum of the outputs of all of its principal dendrites, which means that each principal dendrite of a cell influences – via its contribution to the cell's output – the strengths of synaptic connections on *all* of the cell's dendrites. Because of the Hebbian aspect of synaptic plasticity (Singer 1995; Paulsen and Sejnowski, 2000), the effect of this influence will be for each principal dendrite to drive the other dendrites of the same cell to learn to behave the way it does. In other words, each principal dendrite will *teach* the other dendrites of the same cell to behave the way it does, and it will also learn from them. Through this mutual teaching and learning, all the dendrites in a cell should learn to produce *correlated* outputs, or 'speak with a single voice.'

The fourth consideration is that the different principal dendrites of the same pyramidal cell receive different connections and consequently are exposed to different sources of information about the environment. For example, it is very unlikely that any given axon making a synaptic contact on one dendrite will also have synapses on *all* the other principal dendrites of the same cell (Schuz, 1992; Thomson and Deuchars, 1994). It can, and frequently does, have synapses on more than one principal dendrite, but not on all of them (Deuchars *et al.*, 1994; Markram *et al.*, 1997a). Furthermore, connections coming from different sources have prominent tendencies to terminate on different dendrites. For example, neighboring pyramidal cells make synaptic connections preferentially on the basal dendrites (Markram *et al.*, 1997a), whereas more distant cells, including ones several millimeters away in the same cortical area, terminate preferentially on the apical dendrites (McGuire *et al.*, 1991). Another system of connections, the feedback connections from higher cortical areas, terminate preferentially on yet another part of the dendritic tree, i.e., on the terminal tuft of the apical dendrite, located in layer 1 (Cauller, 1995). Yet another source of differences in inputs to different principal dendrites of the same cell is cortical topographic organization. Across a cortical area, the functional properties of cells change very quickly: even adjacent neurons carry in common less than 20% of stimulus-related information (Gawne *et al.*, 1996; for a review, see Favorov and Kelly, 1996). Basal principal dendrites extend in all directions away from the soma and thus spread into functionally different cortical domains. As a result, these dendrites sample different loci of the cortical topographic map and are exposed to different aspects of the cortical representation of the environment (Malach, 1994).

Referring back to the third consideration: if the 5–8 principal dendrites of the same pyramidal cell have to learn to produce correlated outputs, but are exposed to different sources of information about the environment, then *they will have to tune to different things in the environment that are mutually predictive*. Each principal dendrite in a given pyramidal cell will have to learn to predict, on the basis of the synaptically transmitted information available to *it*, what the cell's *other* principal dendrites are doing in response to *their* inputs. Thus we find that the principal dendrites of pyramidal cells are set up for finding correlated functions: they are given access to different sensory inputs and they are made to find such functions over these inputs that would behave identically, or at least very similarly.

The final consideration concerns the ability of principal dendrites to teach each other nonlinear functions. Cortical synaptic plasticity appears to be Hebbian (e.g., Singer, 1995; Paulsen and Sejnowski, 2000), but the current mathematical formulations of this rule (such as, for example, Sejnowski's, 1977, covariance rule, or the BCM rule of Bienenstock *et al*., 1982, or Grossberg's, 1974, star rules, etc.) all are very weak at learning input-to-output transfer functions, and basically incapable of learning the most useful, *nonlinear* functions (e.g., Willshaw and Dayan, 1990; Hancock *et al*., 1991; Phillips and Singer, 1997). This limitation would greatly reduce the ability of pyramidal cells to find functions that would correlate with each other. However, we believe the problem is not that the synaptic rule used in the neocortex is weak, but that our current theoretical formulations do not capture it adequately. The most recent experimental findings indeed demonstrate that neocortical synaptic plasticity is a much more complex, multifactorial phenomenon than is reflected in current formalisms. (This is a very active field of neurobiological research now, with a continuous stream of new important experimental findings; for just small sampling, with an emphasis on reviews, see Spruston *et al*., 1995; Magee and Johnston, 1997; Markram *et al*., 1997b, 1998; Stuart *et al*., 1997; Svoboda *et al*., 1997; Johnston *et al*., 1999; Malinow *et al*., 2000; Paulsen and Sejnowski, 2000.) Thus, it is quite likely that the real synaptic rule will turn out to have greater capabilities than those for which it is currently given credit. Based on the functional considerations discussed in this paper, we expect that the synaptic rule will turn out to endow principal dendrites with the capacity to teach one another simple or even moderately complex nonlinear input-to-output transfer functions. This is something that ought to be explored experimentally.

All the considerations reviewed above strongly point to individual pyramidal cells – and more specifically, their principal dendrites – as the most likely candidates for performing the task of finding correlated functions over different sets of environmental variables. This is in keeping with Phillips and Singer's (1997) proposal. It remains to be seen how complex are the functions that the principal dendrites of pyramidal cells are capable of teaching each other. We predict that they should be capable of teaching each other to perform nonlinear functions on their synaptic inputs. In this regard we deviate from Phillips and Singer. In another substantial deviation from their idea, we propose that the inputs to a cell are divided into functional sets (for purposes of mutual teaching) by virtue of being located on different principal dendrites of a cell, and not necessarily by whether they are contextual or come from the cell's receptive field.

3. The SINBAD Model of the Pyramidal Cell

We have developed a model of the neocortical pyramidal cell that embodies our idea of how it functions, suggested by the considerations in the previous section. In the version of the model that we describe in this paper, a pyramidal cell has two principal dendrites. Real pyramidal cells have 5–8 principal dendrites, but two dendrites are sufficient for presenting our approach here.

Our modeling approach was shaped by the following consideration. Since we predict that principal dendrites should be able to teach each other nonlinear input-to-output transfer functions, but at this time we do not know its biophysical mechanism, we cannot rely on a standard compartmental modeling approach to render the dendrites (Segev *et al.*, 1989). Instead, we represent each principal dendrite by an error-backpropagating network (Rumelhart *et al.*, 1986). The reason for this choice is that according to our proposal, each principal dendrite should be capable of learning, under supervision by the other dendrites, functions over its inputs that might be nonlinear. In this regard dendrites resemble artificial backpropagation nets. Thus in our model we substitute dendrites with what we believe are their functional approximations.

Thus, representing a principal dendrite by a backprop net, we model a pyramidal cell as a pair of error backpropagating networks whose outputs are added together to produce the cell's output. The cell's output is also used as the training signal for each dendrite. We designate the class of pyramidal cell models to which this one belongs by the acronym 'SINBAD.' This stands for a ***S**et of **IN**teracting* ***BA**ckpropagating **D**endrites*, where "backpropagating" refers to the signal that gets sent back along the dendrites from the soma, affecting synaptic plasticity. The existence of such a signal has been confirmed experimentally (see our "third consideration" in the previous section), though its functional implications remain obscure. In the model presented in this paper, we interpret this signal as the error signal in the standard error backpropagation algorithm; i.e. it reflects the difference between the output of the dendrite and the (scaled) total output of the cell. The signal backpropagated in real dendrites may correspond to this error signal (an idea that should be explored experimentally), or perhaps to something else that gives the dendrites somewhat lesser abilities in terms of the complexity of the functions they are able to teach each other. Our SINBAD idea yields a useful organizing principle by which to guide studies of the content of this and other possible backpropagating

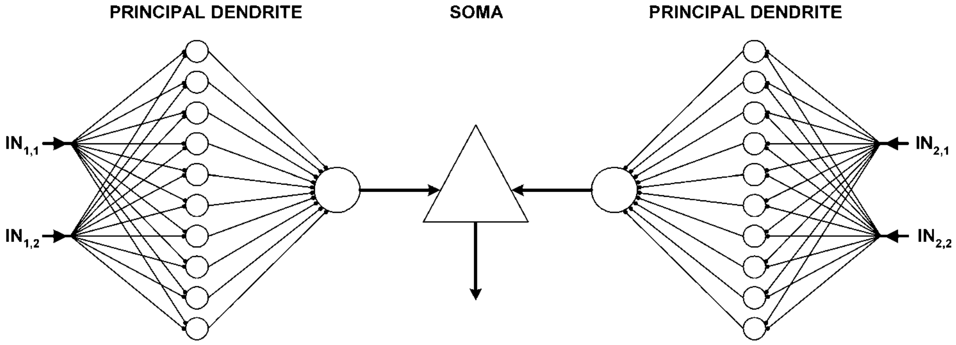


Figure 1. The SINBAD model of a pyramidal cell with two principal dendrites connected to the soma (shown as a triangle). Each principal dendrite is modeled as an error backpropagation network with one output unit, a single layer of ten hidden units, and two input channels.

signals. We can justify our functional level simulation using error backpropagation as an inference to the best explanation: it best explains our psychological abilities, and is consistent with our current knowledge of dendritic physiology.

The design of the cell is shown in Figure 1. In this particular demonstration, each principal dendrite is modeled as a standard backpropagation network (Rumelhart *et al.*, 1986) with one layer of 10 hidden units and one output unit. The hidden units of each dendrite are given input connections from two other cells, or *input channels*, that carry information about environmental conditions. In this demonstration the two principal dendrites receive connections from two different pairs of input channels, $IN_{1,1}$ and $IN_{1,2}$ vs. $IN_{2,1}$ and $IN_{2,2}$.

The activity of a hidden unit *h* in dendrite *d* is computed as a sigmoid function of the activities of its two input channels:

$$H_{d,h} = \tanh(w_{d,1,h} \cdot IN_{d,1} + w_{d,2,h} \cdot IN_{d,2}), \tag{1}$$

where $w_{d,1,h}$ and $w_{d,2,h}$ are the weights of the connections of the two input channels *d,1* and *d,2* onto the hidden unit *h* of dendrite *d.*

The activity of the output unit, i.e. the output of dendrite *d*, is:

$$D_d = \sum_{h=1}^{10} w_{d,h} \cdot H_{d,h}, \tag{2}$$

where $w_{d,h}$ is the weight of the connection from the hidden unit *d,h* to the output unit.

The outputs of the two dendrites are summated to produce the somal input *SOM* = D_1 + D_2. The output of the entire cell is:

$$OUT = \tanh(\gamma \cdot SOM), \tag{3}$$

where γ is a variable that adjusts the cell's output by whether the somal input is greater or smaller than the average somal input. Specifically, $\gamma = 1.2$ if $|SOM| > |\overline{SOM}|$ and $\gamma = 0.8$ if $|SOM| \leq |\overline{SOM}|$. This adjustment drives the cell to expand the dynamic range of its output values. For demonstrating our idea, it is not important here that, in deviation from biological realism, the cell's output can be either positive or negative.

In this demonstration, the sensory environment of the cell was conceived to involve a single orderly entity, object *X*, that can be either present or absent in any given situation (i. e., $X = 1$ or $X = 0$). Object *X* manifests itself in two functions of environmental conditions, each of which is represented by the activity (0 or 1) of one of the input channels. Specifically,

$$X = IN_{1,1} \textit{ exclusive-OR } IN_{1,2} \qquad \text{and} \tag{4}$$

$$X = IN_{2,1} \textit{ exclusive-OR } IN_{2,2} \tag{5}$$

That is, when present, object *X* can reveal itself by activating either input channel $IN_{1,1}$ or $IN_{1,2}$, but not both of them, and also by activating either channel $IN_{2,1}$ or $IN_{2,2}$, but not both of them.

There are eight possible patterns of input channel activities that satisfy Equations 4 and 5, and they define the entire repertoire of environmental situations distinguished by the input channels. All these patterns are shown in Figure 2B. An inspection of these patterns will reveal that the sensory environment is orderly, but this order is not reflected in activities of single channels, and is only apparent at the level of specific (*exclusive-OR*) functions of two pairs of channels. Can the cell discover that the combining functions defined by Equations 4 and 5 are predictive of each other and also informative of their underlying causal source, object *X*? Note that the existence of object *X* is not indicated to the cell in any direct way, but it is only hinted at indirectly by the regularities hidden in the input patterns.

After initially setting all the adjustable connections to randomly chosen strengths *w*'s, the cell was activated with a randomly chosen sequence of the eight training input patterns. The connections were adjusted according to the error back-propagation algorithm after each input pattern presentation. Specifically, the error signals δ_d were first computed for the two dendrites as:

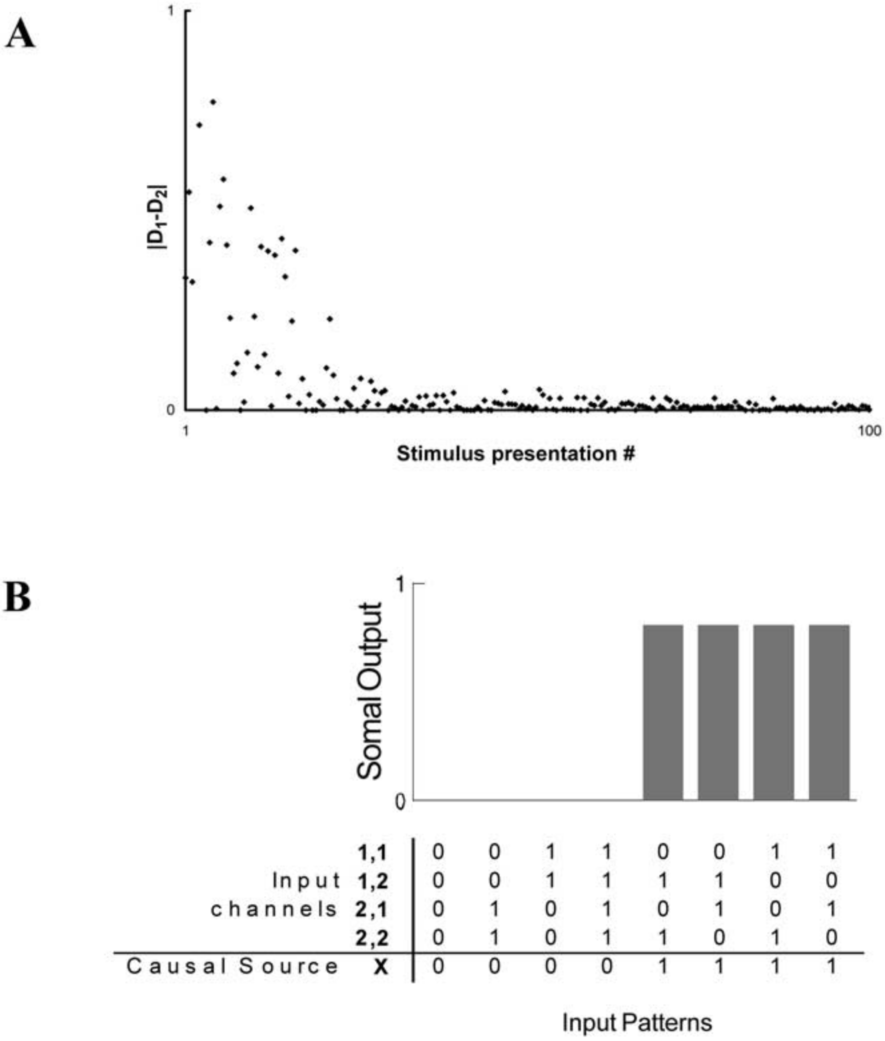
$$\delta_d = (OUT - EST_d) \cdot EST'_d = (OUT - EST_d) \cdot (1 - EST_d^2), \tag{6}$$

where EST_d is the dendrite *d*'s estimate of the cell's output *OUT*. It was computed as:

$$EST_d = \tanh(2 \cdot D_d). \tag{7}$$

For the hidden units, δ was backpropagated as:

$$\delta_{d,h} = \delta_d \cdot w_{d,h} \cdot H'_{d,h}. \tag{8}$$



Input channels	1,1	0	0	1	1	0	0	1	1
	1,2	0	0	1	1	1	1	0	0
	2,1	0	1	0	1	0	1	0	1
	2,2	0	1	0	1	1	0	1	0
Causal Source	X	0	0	0	0	1	1	1	1

Figure 2. Learning performance of the SINBAD model. **A.** Magnitude of the difference in outputs of the two dendrites in response to a random sequence of input patterns. In response to each input pattern, the two dendrites adjusted their connections, showing a rapid learning progress, which was essentially completed by the 20th stimulus presentation. **B.** Output of the cell, *OUT*, in response to the 8 training input patterns. The training patterns are shown in the table below, aligned with the plot. Note the match between the state of the causal source *X* and the cell's output.

Connection weights were adjusted by:

$$\Delta w_{d,i,h} = \mu_i \cdot IN_{d,i} \cdot \delta_{d,h} \quad and \tag{9}$$

$$\Delta w_{d,h} = \mu_h \cdot H_{d,h} \cdot \delta_d, \tag{10}$$

where μ_i and μ_h are learning rate constants for the input and hidden unit connections (μ_i = 6 and μ_h = 0.003 in our simulations).

Before we turn to the results of simulation of this model neuron, it might be helpful to consider the nature and dynamics of the learning task we set up for the two dendrites. Being a backpropagation network, each dendrite can, in principle, learn a large variety of input-to-output transfer functions. Its actual choice will be dictated by the teaching signal, which comes from the other dendrite. But that dendrite also has a large choice of possible transfer functions, and it itself relies on the first dendrite for its own guidance. Thus, the two dendrites will teach each other how to respond to input patterns while continuously changing their own behaviours and their own teaching signals. Such a teaching/learning process will continue until the two dendrites discover such transfer functions that will enable them to have identical responses to their *different* co-present inputs. In other words, the process of connection strength adjustments will continue until each dendrite will learn to predict – on the basis of its own inputs – the responses of the other dendrite to its inputs.

Of course, if the input patterns applied to one dendrite do not relate in any way to the input patterns applied at the same time to the other dendrite, but are accidental in their co-occurrence, then it will be impossible for the dendrites to discover any matching transfer functions, and the process of connection strength adjustments will continue indefinitely. On the other hand, if there is some consistent, and therefore predictable, relationship between co-occurring patterns of the input to the two dendrites, then the dendrites might be able to discover transfer functions predictive of each other's outputs, thus associating a pair of complex conditions. Success will not be guaranteed, but will depend on the complexity of the orderly relations between the two sets of inputs. For our demonstration here we chose an intermediate level of complexity of the orderly relationship between input channels $IN_{1,1}$–$IN_{1,2}$ and $IN_{2,1}$–$IN_{2,2}$; as described above (Equations 4–5), this relationship is not discernable at the level of individual channels, but only at the level of their *exclusive-OR* logical functions.

Figure 2 shows the progress and the results of the two dendrites teaching each other how to respond to input patterns. The most pressing question is: will the two dendrites learn to predict each other's responses to the input patterns? To answer this question, the difference in the output activities of the two dendrites (i. e., $|D_1 - D_2|$) was plotted in Figure 2A as a function of each successive input pattern presentation. This plot shows that the two dendrites discovered very quickly how to respond to their inputs so that they would produce identical outputs. Exposure to a random sequence of fewer than 20 input patterns was sufficient for the two dendrites to discover a way to produce nearly identical responses to their *different* inputs.

The dendrites' learning success raises the following question: what did the dendrites discover about the environment that enabled them to predict each other's responses? To answer this question, we need to examine responses of the whole cell to the entire repertoire of eight possible input patterns. These responses (plotted in Figure 2B) were obtained after the cell was exposed to a random sequence of 100 input patterns, by which time the two dendrites had already learned to respond virtually identically (see Figure 2A).

Judging by the cell's responses plotted in Figure 2B, the first dendrite learned to respond to the *exclusive-OR* function of its input channels $IN_{1,1}$ and $IN_{1,2}$, while the second dendrite learned to respond to the *exclusive-OR* function of its input channels $IN_{2,1}$–$IN_{2,2}$. Because of the way we set up the sensory environment (Equations 4–5), this pair of transfer functions is the only pair that produces identical outputs, and that is why the dendrites chose them. Thus, Figure 2B shows that the two dendrites successfully identified two correlated nonlinear functions of their inputs; we call these 'order-revealing functions' (as opposed to arbitrary functions).

The functional significance of the outcome of the dendrites' learning goes beyond the discovery of two correlated, order-revealing functions of environmental conditions; it identified the causal source of this order. This causal source is object *X*: the reason why the two functions of conditions are predictive of each other is because they originate from the same source, object *X*. As Figure 2B shows, the cell's output is accurately indicative of the presence and absence of object *X*. The cell, in effect, discovered the existence in the environment of object *X*, in that its activity carries explicit information about the presence of object *X*.

In conclusion, this modeling exercise shows how (1) by teaching each other the dendrites in a cell can tune to different but correlated, order-revealing functions of environmental conditions, and (2) the cell, as a whole, can learn to recognize the source of those correlated functions, an orderly feature in its sensory environment. The first ability implements association of functions over conditions, in contrast with Hume's simple association. The second ability (which is really just the flipside of the first) implements a process of representation acquisition, to replace Hume's abstraction.

4. How to Divide the Connections Among the Dendrites?

So far we have described the principle by which pyramidal cells can discover regularities in their sensory environment. Now, for any practical implementation of this principle, we must also address the question of how to distribute the input channels among the principal dendrites of a cell so that the right sets of channels will go to the right dendrites. In the previous modeling exercise we skirted this issue by assigning the four input channels to the two dendrites in pairs that we knew were the right ones for the environment we set up there. But if we did not

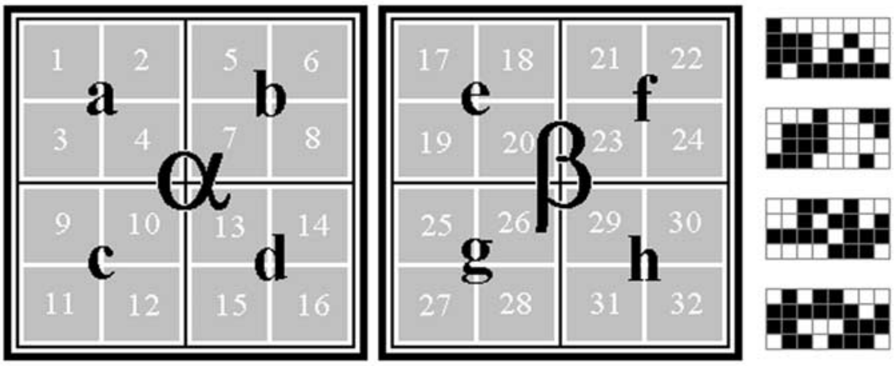


Figure 3. The orderly structure of the model environment. Shown as small gray squares are 32 elementary environmental conditions (binary variables); the state of each of these is given directly by the activity of an input channel, i.e. 0 or 1. (A tiny sample of the possible combinations can be seen in a column on the right, where shaded squares represent active channels.) These elementary conditions are organized into 8 sets that indicate, according to some function, environmental properties *a–h*. These 8 properties, in turn, are organized into 2 sets that indicate objects α and β. The rules by which the status of higher-order conditions are indicated by lower-order conditions are specified in the text.

know the orderly organization of the environment in advance, then we would not know how to divide the four channels between the two dendrites. Another concern is that, unlike in the previous exercise, the environment possesses not just one, but many different orderly features and of various levels of complexity. How can the cells discover as many of these orderly features as possible?

To address these issues, we will start with another simple modeling demonstration. For this demonstration we created a more complex sensory environment. This environment is characterized by 32 binary parameters, or *elementary conditions*. These might be, for example, 32 sensory channels through which some hypothetical animal obtains information about the state of its surroundings. The environment is orderly, involving two levels of regularities.

The orderly structure of the modeled environment is shown in Figure 3. The environment was set up to have two complex entities, called objects α and β. Object α is indicated by environmental properties *a* and *b*, let's say α = (*a exclusive-OR b*). That is, object α manifests itself as *a* or *b*, whereas together *a* and *b* do not indicate α, but form an accidental combination. In addition, α is indicated by a function of properties *c* and *d*, let's say α = (*c exclusive-OR d*). It might be, for example, that α has two sides, one side characterized by either *a* or *b*, and the other by either *c* or *d*. Object β is organized according to the same plan as object α: β = (*e exclusive-OR f*), and β = (*g exclusive-OR h*).

In another layer of complexity, environmental properties *a, b, c, d, e, f, g, h* are in turn indicated by elementary environmental conditions *1* through *32*. All of these properties were given the same organization:

a = (*1 exclusive-OR 2*) = (*3 AND 4*) *b* = (*5 exclusive-OR 6*) = (*7 AND 8*)
c = (*9 exclusive-OR 10*) = (*11 AND 12*) *d* = (*13 exclusive-OR 14*) = (*15 AND 16*)
e = (*17 exclusive-OR 18*) = (*19 AND 20*) *f* = (*21 exclusive-OR 22*) = (*23 AND 24*)
g = (*25 exclusive-OR 26*) = (*27 AND 28*) *h* = (*29 exclusive-OR 30*) = (*31 AND 32*)

Thus, the modeled environment possesses 32 elementary conditions, 8 first-order regularities (due to properties *a–h*), and 2 second-order regularities (due to objects α and β).

We presented this environment to the model used in the previous demonstration, with a few modifications. The first modification is that this cell now has 32 input channels, each carrying information about one of the 32 elementary environmental conditions. Unlike the previous exercise, these connections are divided randomly between the two dendrites of the cell. Each input channel is assigned at random to either one or the other dendrite, but not to both of them. Also, the connection weights of the input channels are set initially to zero, except for four randomly chosen channels on each dendrite. For these four connections, their initial connection weights are chosen at random. Thus, unlike the previous exercise, here we do not take advantage of our knowledge of which input channels should go together.

In another modification of the initial design, the number of hidden units in each dendrite is increased to 40. Otherwise, the activities of the hidden units $H_{d,h}$ and the dendrite's output unit D_d are computed as described in Equations 1 and 2. The cell's output *OUT*, error signal δ_d, error backpropagation $\delta_{d,h}$, and dendritic connection *w* adjustments are computed as before, according to Equations 3, 6–10.

Figure 4 shows the result of one simulation run during which 10,000 input patterns were presented to the cell in a random sequence, taken from the 2^{22} patterns that can be encountered in the environment defined above. To see whether the two dendrites of the cell learned to produce similar outputs in response to their co-present input patterns, the correlation coefficient between outputs of the two dendrites was computed across 100 successive input pattern presentations. In Figure 4 the value of this correlation coefficient is plotted as a function of time from the start of the training period. This plot shows that in the beginning the two dendrites correlated poorly in their outputs, but after some initially unsuccessful search the dendrites discovered a way to produce very similar outputs. This means that the two dendrites discovered some orderly relationship in the environment, which enabled them to predict each other's responses to input patterns.

The environment has a number of orderly features, listed above, and which of them were discovered by the cell was determined by the distribution of the input channels on the two dendrites. Because the channels were distributed between the dendrites randomly, there is no guarantee that a set of channels engaged in a given order-revealing function (e.g., channels *1* and *2*, as an *exclusive-OR* function, reporting on the status of property *a*) were placed on the same dendrite, or that predictive pairs of channels were placed on the opposite dendrites (e.g., channels *1* and *2* on one dendrite and channels *3* and *4* on the other dendrite). As a result, the

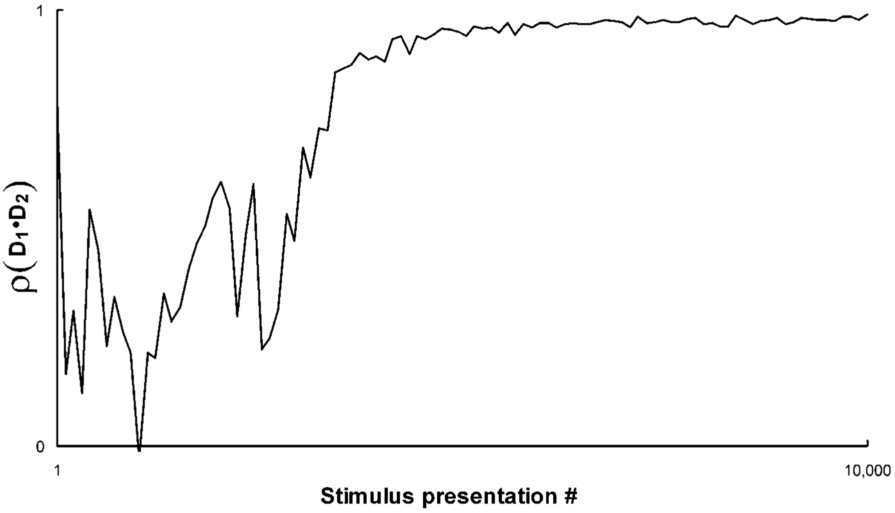


Figure 4. Learning performance of the SINBAD model. The correlation coefficient, ρ, between the outputs of the two dendrites is plotted as a function of time from the start of the training period.

particular distribution of input channels on the dendrites might make it impossible for the cell to discover a particular environmental regularity.

There are a number of ways a cortical network is likely to deal with this limitation. The simplest, most naive way is to have many pyramidal cells. Some of them will, by accident, have an appropriate distribution of input channels on their dendrites and thus they will be in a position to discover that orderly environmental feature. With a sufficient number of cells, the network will be able to discover all the orderly features.

To illustrate this idea, we expanded the model to have 32 cells identical to the one used in the last exercise. They differ from each other only in the distribution of input channels on their dendrites, which was assigned randomly. For simplicity, the 32 cells were run in parallel, without interactions among them. To evaluate the learning outcome in this network, we also set up three additional cells (or more accurately, one of the principal dendrites of each of three additional cells that are supposed to reside in a higher cortical area). The design is shown in Figure 5. Our aim here is to compare the ability of a 'test' dendrite to learn a particular behaviour, given either raw information provided by the input channels, or the information provided by the 32 cells (which presumably transformed the raw input patterns into a new form in which some of the orderly environmental features *a–h* are represented more explicitly).

One test dendrite was trained to respond to the presence of object α, another to object β, and the third to $\Omega = (\alpha$ *exclusive-OR* $\beta)$. These dendrites might be viewed as belonging to cells that, for some unspecified reasons, are driven by their other dendrites to respond to α, β, or Ω.

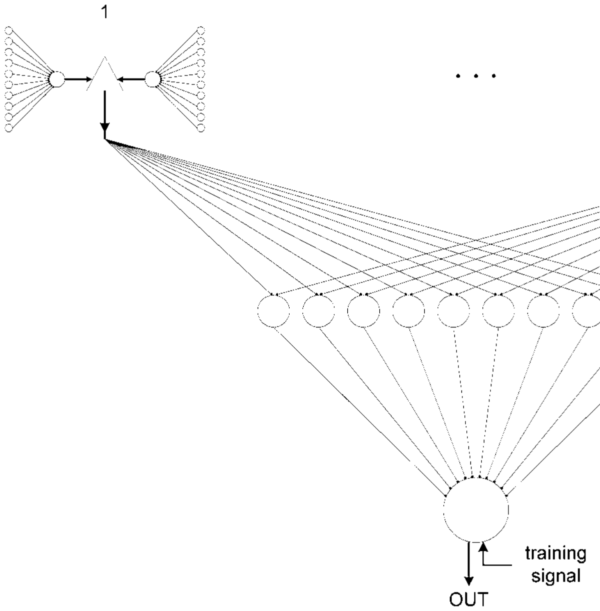


Figure 5. A layer of 32 SINBAD cells whose outputs are connected to a test dendrite, which is modeled as a backpropagation network with one output unit and a single layer of hidden units.

The three test dendrites are modeled the same way as the dendrites of the 32 cells, except that the test dendrite receives connections from all 32 cells (or 32 input channels, in the raw input test) and its output is passed through a sigmoid function:

$$D = \tanh(\sum_{h=1}^{40} w_h \cdot H_h). \tag{11}$$

The error signal is computed as:

$$\delta = (TR - D) \cdot D', \tag{12}$$

where *TR* is the training signal (the status of α, β, or Ω). Effectively, these test dendrites are modeled as standard three-layer backpropagation networks.

To establish a basis for comparison, we first trained the test dendrites on raw sensory information, by using input channels $IN_1 - IN_{32}$ for their inputs. The results are shown in Figure 6A, for the test dendrite trained on object α (see the plot on the left) and for the test dendrite trained on object Ω (right plot). Each plot shows the *error* (which is computed as |*TR* – *D*|) as a function of time since the start of the training period. As the left plot shows, the α dendrite gradually succeeded in learning to recognize object α. The right plot, on the other hand, shows that the Ω

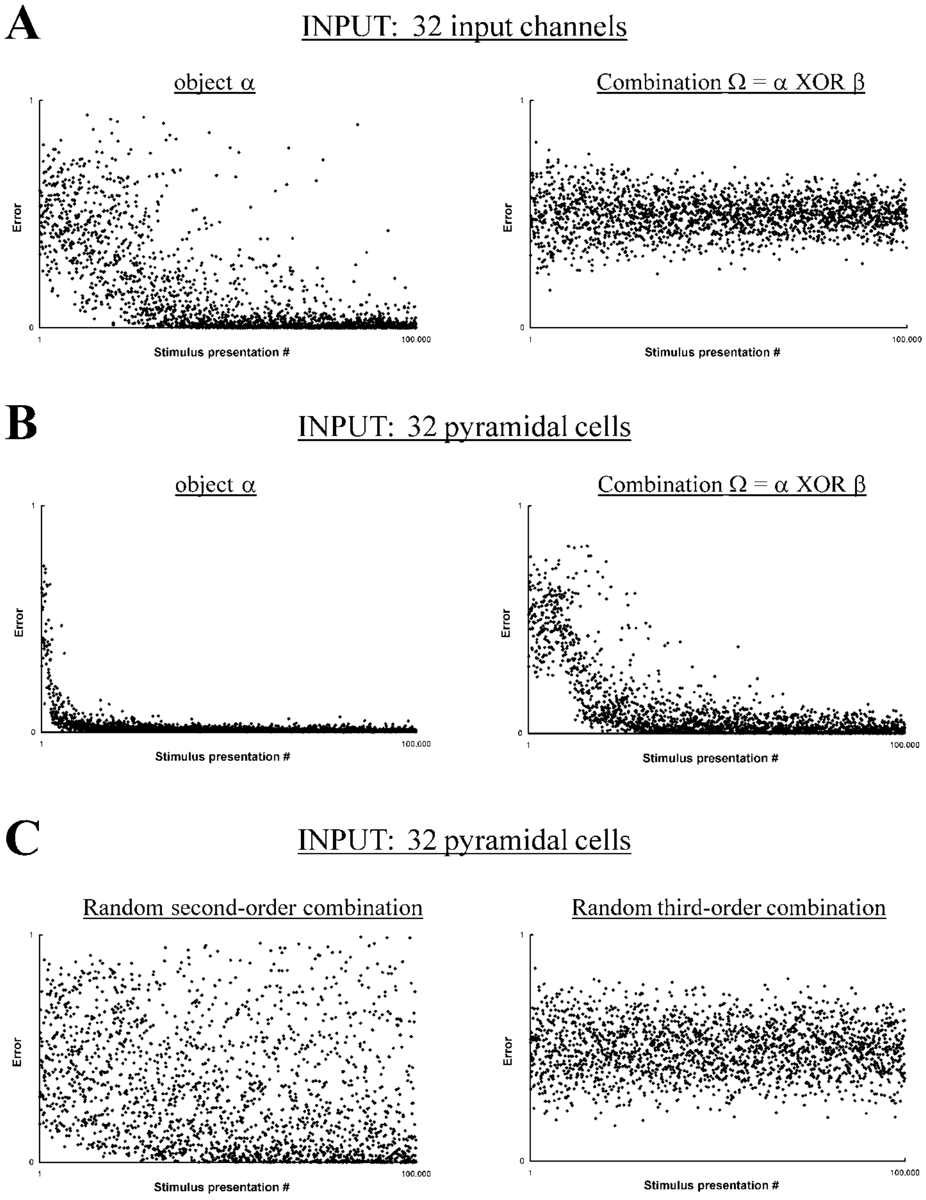


Figure 6. Learning performance of two test dendrites (left and right columns) in response to a random sequence of 100,000 input patterns. The error plotted is the magnitude of the difference between the training signal *TR* and the test dendrite's output *D*, showing how well the dendrite predicted the training signal. Plotted are responses only to every 100th input pattern. **A.** Two test dendrites, trained to recognize object α (left) and object Ω (indicated by a function of objects α and β), received their inputs directly from the 32 input channels, rather than from the 32 SINBAD cells. **B.** The α and Ω test dendrites received their inputs from the 32 SINBAD cells, rather than directly from the input channels. **C.** Two test dendrites, with their inputs coming from the 32 SINBAD cells, were trained to recognize two randomly defined second- and third-order functions of elementary environmental conditions.

dendrite failed to learn to recognize object Ω. This failure is not surprising, since Ω is correlated with a third-order nonlinear function, and backpropagation networks have great difficulty in learning such complex functions (Clark and Thornton, 1997).

We next trained the test dendrites on sensory information processed by the 32 SINBAD cells: outputs of the 32 cells were used as inputs to the test dendrites. The results are shown in Figure 6B, again for the α dendrite on the left and for the Ω dendrite on the right. As these two plots show, the learning performance of the test dendrites improved dramatically. The α dendrite learned to recognize object α orders of magnitude faster than when it learned from the raw input channels. And the Ω dendrite – which before could not learn at all – now did learn to recognize object Ω. What is particularly impressive is that the test dendrites learned to respond correctly to the input patterns after being presented with only a small fraction ($<0.5\%$) of all possible input patterns. That is, the test dendrites showed perfect generalization abilities; they discovered the logic underlying the relationship between the input patterns and the behaviours on which they were trained. The performance of the Ω dendrite is especially impressive as it was able to learn a very challenging, third-order function of input channels.

What is the nature of the information preprocessing, carried out by the layer of the 32 pyramidal cells, that improved so dramatically the learning abilities of the test dendrites? During the period of the network's learning, pairs of dendrites in each cell taught each other to tune to correlated functions of conditions. Significantly, these functions of conditions were due to properties *a–h*. With the dendrites tuned to individual order-revealing functions, the entire cells tuned to recognize the presence and absence of those properties *a–h*. As we discussed above, due to random assignment of input channels on the dendrites, different cells tuned to different properties among *a* through *h*, but as a group, the 32 cells were likely to discover all 8 of them. In this way, the information about the status of *a–h* was brought to the surface in the 32 cells' outputs.

With properties *a–h* represented directly by individual pyramidal cells, the learning task for the test dendrites was simplified: for the α dendrite, for example, the task was changed from that of learning a second-order nonlinear function of input channels *1–32*, to the much easier one of learning a first-order function of properties *a–h*. For the Ω dendrite, its task was simplified from learning a third-order function to a second-order one.

If this interpretation of the Figure 6B results is correct, then the improvement in learning by test dendrites should be limited only to *order-revealing* functions of environmental conditions. If we were to ask the test dendrites to learn some second- or third-order functions of input channels that do not involve properties *a–h*, but are simply arbitrary in their composition, then information preprocessing by the 32 cells should not be of any help, since the cells will not make explicit the 'building blocks' of such arbitrary functions. To test this prediction, we trained the test dendrites on the outputs of the 32 cells, but using different training signals. The α and β dendrites were trained to recognize second-order functions of elementary environmental conditions (just as objects α and β are indicated by such second-order functions), but these new functions had arbitrary compositions, not involving environmental properties *a–h*. Analogously, the Ω dendrite was trained to recognize an arbitrary third-order function.

The results are shown in Figure 6C, and they confirm our expectation. Note especially that the learning performance of the dendrite trained on a second-order function (shown in the left plot) is much worse than when that dendrite was trained to recognize a comparable function (i.e., object α) from the raw information provided by input channels (see left plot in Figure 6A). This deterioration of learning performance is not surprising, suggesting that the 32 cells made some functions of environmental conditions – the order-revealing ones – more explicit, in part, by filtering out arbitrary functions.

In summary, this modeling demonstration shows that SINBAD cells discover regularities in the environment remarkably easily, even when the input connections are distributed among dendrites at random. With their dendrites tuning to correlated functions of environmental conditions, cells learn to recognize the orderly features of the environment that cause these functions to obtain. Obviously, pyramidal cells in primary sensory areas are exposed only to raw peripheral inputs and thus will only be able to tune to stimulus features engaged in *local* orderly relations. However, if the outputs of those cells are fed to another layer of pyramidal cells (e.g., primary visual cortical area V1 projecting to V2), these cells will be able to use the features discovered by the first layer to tune to the next generation of environmental features. We demonstrated this phenomenon with the α and Ω test dendrites. Such incremental elaboration is possible because complex orderly features of the environment are indicated by sets of simpler orderly features, and ultimately by receptor activities. Thus, for example, different types of situations (involving interacting objects) are composed of various types of objects and their states and relations, which have stereotypical surfaces and figures etc., which are in turn indicated by lines, edges, and textures, which are complex invariants to be found in patterns of receptor activities.

We should point out that although we did not take advantage of them here, a number of biologically realistic means are readily available by which the arrangement of input connections on the dendrites can be optimized. Among them are: (1) cortical topographic maps constrain the choice of inputs available to a cell to those most likely to have orderly relations, (2) connections on dendrites can be rearranged by trial and error, (3) lateral inhibitory connections on cells' somata can diversify functional properties of neighboring cells, (4) trophic signals from postsynaptic cells can be used as a measure of usefulness of a cell's output and drive cells to tune to the most significant regularities. A detailed discussion of these means is, however, beyond the scope of this paper; it can be found in Favorov *et al*. (2001).

5. Representation Acquisition in SINBAD Networks

If the principal dendrites of a cell compute correlated functions, the cell as a whole will tend to tune to the source of those correlations, if there is one. Of course, non-accidental correlations (the only ones that are useful) will by definition have an *explanation*, and thus a source in some sense. In the simulations presented in the previous section, properties a to h, and objects α and Ω were examples of sources of correlation. What kinds of sources of correlation might be discovered by SINBAD cells? Here are some major ones:

- *Linking or Common Causes* – the learned feature might be something invisible, but detectable indirectly from the things that control it, and/or its effects on yet other things. For example, Mendel discovered the existence of hereditary factors, genes (linking causes), by observing orderly relations among the traits of parents and their offspring, and medical researchers infer the existence of a pathogen (common cause) from multiple coincident symptoms of the disease that it causes. At the other end of the scale of abstraction, the property of surface reflectance is an invariant that links illumination conditions and the light that enters the eye in various complex ways. Perhaps the representation of surface color is an imperfect attempt of SINBAD cells to associate functions of reflected light and illumination.
- *Structures* – one set of variables might describe a distinct combination of features of one part of a stereotypical entity (kind) or a unique-but-persistent entity (individual), and another set of conditions might describe a distinct combination of features of another part of it. For example, a head and a body go together as two parts of the same natural kind, a human being, or a particular distinct head and a particular distinct body go together as two parts of the same individual, your particular friend.
- *Functional kinds* – one set of variables might describe properties intrinsic to the entity itself, and another set will describe its effects. For example, an ape might acquire the concept of a club by observing a relationship between tree branches of certain shapes and sizes and their effects on its enemies.
- *Mixtures of these* – kinds, broadly construed, can have a mixture of intrinsic, causal, structural, and functional properties in common. All of these will be mutually correlated in various combinations.

Stated generally, under our proposal a pyramidal cell can be expected to tune to a set of related functions of environmental conditions grounded in an orderly feature of the environment. Especially in the case where there are *multiple* correlations, it seems likely that their source will be some re-identifiable entity – an individual, a kind of object or event, or a property – that explains the correlation of the functions a cell's dendrites were able to discover. The process of a SINBAD cell's tuning to a source of correlation is the mechanism we propose to replace Hume's abstraction.

Earlier, we described the problem that Hume's process of abstraction faced, assuming it is meant to be part of a naturalistic theory of the mind. The problem was this. Naturalism numbers among its commitments a mechanistic explanation of rational processes. If a theory explains a mental occurrence by saying "it stands to reason," and says nothing further about how this rational process occurs, the theory thereby fails to be naturalistic. Hume's theory of abstraction fails to be naturalistic in just this way. Hume says that where there is similarity, an abstract representation of that similarity is created. But there is far too much similarity around, especially if one includes higher-order similarity (which one must, both to make higher-order regularities available for prediction, and to accurately depict human cognition). We know that the mind creates new representations of those features that are predictively related to others, but the only explanation Hume can offer for why it is *these* features that are abstracted out is an essentially rational one: the mind *chooses* to abstract out those features because they are the ones that are predictively useful.[3]

On the SINBAD model, the problem disappears. The process of representation acquisition is guided by predictive utility, and there is a fully mechanistic explanation for this. In fact, the process of representation acquisition and the process of associating predictively related features are two sides of the same coin. The principal dendrites on a SINBAD cell associate correlated functions. Those correlated functions have an explanation in the environment, the source of the correlation. As its dendrites discover their correlated functions, the cell as a whole tunes to the source of correlation. Exposed only to 'surface information' (corresponding to Hume's raw sensations), a SINBAD network creates a new representation of the entity that unifies that surface information. Thus only features that are predictively related to others are "abstracted out."

In the final simulation, properties *a–h* and objects α, β, and Ω were sources of correlated functions of elementary conditions. Though the SINBAD cells are exposed only to information about the elementary conditions, some of them tune to *a–h*, α, β, and Ω. The network acquires representations of these properties and objects because they are central in a network of predictive relations. In our current work (Favorov and Ryder, in preparation), we make use of a concrete example, the kitchen sink, which offers a more intuitive illustration of representation acquisition in SINBAD networks. Given only inputs about the sink knobs (their turning directions, positions, and which is hot and which cold) and water flow at the tap (instantaneous flow rate and temperature), the network acquires representations of two predictively central, linking variables: the flows of water in the hot and cold pipes, i.e. the water flows directly controlled by the knobs. These analog linking variables are hidden from the network, that is to say it receives no inputs carrying information about the water flows in the hot and cold pipes. Despite this limitation, SINBAD cells in the network tune to these hidden vari-

[3] One might charitably suggest that Hume could offer a 'trial and error' model, where if a representation of a predictively useless abstract feature is created, it is abandoned and a new attempt at abstraction made. Construed as a psychological process, this is highly implausible. Construed as a sub-psychological process, it is a crude approximation of what SINBAD actually does.

ables – the network "infers" their existence. This is because the flows in the hot and cold pipes are sources of correlation between functions of the surface variables the network *does* get information about, just like α, β, and Ω are sources of correlated functions of elementary conditions. For the purposes of our simulation above, the functions linking elementary conditions with α, β, and Ω were randomly chosen. In a real system, these functions will reflect real relations amongst the system's variables. In the case of the kitchen sink, the flows in the hot and cold pipes are what causally link the knob positions to flow and temperature at the tap, and vice versa. This endows the flows in the hot and cold pipes with predictive value, which in turn explains why the SINBAD network acquires representations of these variables. Representation acquisition is not accomplished by classifying certain "similar" items together using some arbitrary and limiting measure of similarity. In a SINBAD network, representation acquisition is automatically guided by predictive utility.

It is even more beneficial than it sounds that, in a SINBAD network, new representations are created only[4] for features that are predictively related to others. This is because some correlation is often a sign of more correlation. We hypothesize that the sorts of things that SINBAD cells tune to, namely sources of correlation, tend to be sources of many correlations. The creation of a new representation embodied in a SINBAD cell occurs on the basis of a few correlations discovered by that cell's dendrites. Those correlations are explained by a certain source, to which that cell has tuned. If that source yields further correlations, the dendrites of the cell will be in an excellent position to discover these further correlations. In effect, the cell will learn more and more ways by which to identify the source, and by the same token, more features that item predicts.[5] That is, if some correlation reliably indicates further correlation, the predictive capacities of a SINBAD network will be that much better. This will indeed be the case if sources of correlation tend to be "substances", in Millikan's sense (1999).

Substances are entities that have *rich inductive potential* (a term Millikan borrows from Gelman and Coley (1991)). Substances possess a number of properties that (more or less reliably) tend to cluster together – cats tend to have fur, meow, move in certain ways, hunt; water has a certain melting and boiling behaviour, is clear, tasteless, etc. Knowing that something is a cat makes available a large amount of knowledge that is not on display in a current encounter: the cat is not now making noise, but you know what noise it would make. In other words, knowing something is a cat allows you to make a number of predictions. It is also a good bet that a property newly discovered to be instanced by a substance on

[4] This does not mean that the SINBAD mechanism entails that categories with little to no predictive utility cannot be represented (e.g. because they are uninstantiated); it does entail that they must be represented compositionally, e.g. one-eyed + one-horned + flying + purple + people eater.

[5] We have confirmed that this effect will be enhanced if the cell has not yet reached a state of *perfect* correlation, either because the functions its dendrites have discovered are not perfectly predictive, or because there is noise in the system (Favorov and Ryder, in preparation).

one encounter will also be instanced on future encounters. To borrow an example from Gelman and Coley, if we find that one cat has a substance called ‘cytosine’ inside, we can reliably induce that other cats also contain cytosine. Being able to re-identify substances makes predictions available for use, and permits the acquisition of further facts that could be used predictively in the future.

The rich inductive potential possessed by substances is not accidental (Millikan, 1999, p. 529). Substances have it for a *reason*. In the case of water, that reason is water’s chemical structure, which explains its surface properties. The hidden explanatory principle in the case of cats is a more complex biological one. Millikan notes (p. 529) that natural kinds are not the only things that yield rich inductive potential – other real kinds, such as artifacts, and individuals do as well. We would suggest that rich inductive potential is even more widespread; it is to be found in event types, interacting elements of dynamical systems, and perceptual properties (like lightness and colour) that can be represented as complex invariants hidden in the inputs to our sensory receptors. All of these things may be thought of as sources of correlation. The important point is this: being able to lock onto or track individuals, real kinds and other sources of correlation that have rich inductive potential enhances an organism’s ability to predict. Re-identifying a kind or individual *as* that kind or individual allows an organism to (a) use the knowledge it has already gained about the kind or individual, and (b) acquire new knowledge upon further encounters with that kind or individual (Millikan, 1999, pp. 531–533). These identifications and re-identifications may be accomplished in any way at all, as long as they are reliable enough to be predictively useful. Typically, an organism will have at its disposal many ways to identify an item, *via* many different subsets of properties/descriptions. It will use the subset that is perceptually available on a particular occasion, using different subsets on different occasions. The larger the number of reliable indicators of a particular source of correlation an organism can collect, the better (p. 532).

As we have already seen, SINBAD networks are particularly suited for discovering and representing sources of correlation, and associating them via the complex functions by which they are related. Therefore, if most of the correlation to be discovered in the environment is due to entities with rich inductive potential, it would be unsurprising if the cortex turned out to be a SINBAD network. Alternatively, even if there is much correlation in the environment that is not concentrated around sources, but it is true that our mental representations tend to be of sources of correlation, this would offer some degree of confirmation to the SINBAD hypothesis. In any case, if an organism possessed a SINBAD network, and it was exposed to sources of correlation, the network would prove to be predictively useful. Exposed to such sources, the network would construct an internal model, consisting of representations of sources of correlation, along with complex associations mirroring the predictive relations into which these sources enter, with each other and with the organism’s sensory apparatus.

6. Concepts in SINBAD Networks

We have suggested replacing Hume's unexplicated notion of abstraction with a rather different and fully naturalistic mechanism of representation acquisition. One important variety of representation acquisition is concept acquisition. The literature on concepts is vast, but theories of concepts (understood as mental representations) can be divided into three main types: descriptionist, referential, and role theories. We believe that there is something right about all three theory types; in fact, they have all tapped into some aspect of the underlying SINBAD mechanism. SINBAD unifies the various theories of concepts, and taken together, they constitute evidence for the truth of the SINBAD model.

The classical descriptionist reduces a concept to a definition (which must not contain the concept to be defined). For example, the concept <bachelor> is reduced to <adult, never married male human>. Each of the component concepts in the definition would receive its own reductive definition, until the original concept is reduced, via many steps, to a stock of undefined primitive concepts. For the phenomenalist, the undefined primitives were 'sense data', roughly equivalent to sensations. While the classical theory was originally intended as a theory of concepts taken as abstract objects, one could adapt the theory for naturalistic psychological purposes by postulating identifying definitions, physically realized, associated with each concept (see Katz, 1972). In the context of SINBAD, this would amount to treating the functions instantiated by a cell's dendrites as a set of descriptions, that together provide a complex definition which is the cell's representational content.[6] Again, some stock of primitive concepts would have to be specified; the purest descriptionist could follow a phenomenalist strategy, a 'naturalized phenomenalism.' The naturalized phenomenalist avoids idealist metaphysical conclusions by replacing the traditional phenomenalist's mental entities – the basic sensory qualities – with physical ones, namely the activities of one's sensory receptors (which carry information about all their possible causes). Cells at the cortical (or thalamic) periphery would receive these 'sense-data', and their content would be reduced directly to them. Cells at the next level up in the cortical hierarchy would have definitions determined by their relations (i.e. the functions embodied in their dendrites) to cells at the periphery. As for the traditional phenomenalist, the reduction to sense-data might proceed by many steps, depending on how far away a cell was from the periphery.

The classical descriptionist theory has fallen on hard times. One main reason for this has been the difficulty of actually producing any concept definitions, even for seemingly obvious ones like the concept of a bachelor (Fodor, 1998). SINBAD

[6] One might try to privilege certain aspects of the functions instantiated in the dendrites as individuating a cell's content; this would require defending some sort of analytic/synthetic distinction against Quine's criticisms (1953). By making all aspects of the functions instantiated in a cell's dendrites relevant to its representational content, the naturalized phenomenalist responds to Quine's challenge by claiming that all is analytic, by contrast with Quine who proposed that all was synthetic.

may explain this difficulty: it may be due to the enormous complexity of the associations embodied in SINBAD cells. The complexity of these associations greatly exceeds Hume's pairwise associations, and even goes beyond Russell's logical constructions (1914/93), since dendrites can instantiate continuous mathematical functions. If the SINBAD theory is correct, there will be *some* complex relation between a concept and the (motor and/or sensory) periphery, embodied in a series of dendritic functions. Whether that relation is helpfully construed as *definitional* is another matter, but there is still something right about the classical descriptionist account.

A more modern descriptionist account of representation is prototype theory and related similarity based accounts (Rosch and Mervis, 1975; Smith and Medin, 1981). The prototype theorist does not believe that concepts can be reduced to definitions. But prototype theory is still descriptionist, since it proposes that concepts may be reduced to a weighted set of features, or a weighted set of similarities to a prototype.[7] Whether something falls under the concept is not a cut and dried matter, as it is for the classical descriptionist; rather it is probabilistic. There is no reason why the dendrites on SINBAD cells could not instantiate probabilistic functions, and it may be this aspect of the underlying mechanism that the modern descriptionist has uncovered.

The descriptionist assumption has been widely questioned in recent philosophy (though something like it is currently enjoying a revival – see Chalmers (1996, Ch. 2) and Jackson (1998)). According to the descriptionist, whatever satisfies the description that defines a concept is thereby a referent of that concept. Some simple examples expose potential problems with this view.

One type of problem arises because one's knowledge is often incomplete, e.g. one might know water only by its surface properties (clear, colourless, tasteless, liquid), not as H_2O, or know Socrates only as "the ancient Greek philosopher guy." Does someone who lacks an identifying description of an individual or kind thereby lack a concept of it? Does someone who is ignorant of chemistry not really have a concept of water, but rather a concept that refers to any liquid that has water's surface properties? Or suppose they have a couple of false beliefs about water, and there is nothing that satisfies their defining description; is their concept *empty*? (If so, they couldn't really have false beliefs about water, they could only conceive of something nonexistent.) Also, each person's incomplete knowledge differs; you probably know quite different things about Aristotle than your neighbour. On the descriptionist assumption, the consequence is that though we may believe we are thinking about (and discussing) the same thing, in fact we are thinking about quite different things until we get our defining descriptions to match precisely. It seems we cannot even dispute facts about Aristotle, or cats, because it is virtually impossible for us to agree on the subject matter.

[7] Note that representation acquisition on these accounts, in contrast with ours, tends to follow Hume's raw similarity based model.

For these and other reasons,[8] the dominant view in philosophy is that concepts, or at least certain types of concepts e.g. concepts of individuals and natural kinds, are not descriptive. 'Referential' theories (as opposed to descriptionist theories) maintain that concepts are not to be individuated (counted as the same concept) according to their associated descriptions, but rather according to their referents (Putnam, 1975 (on linguistic meaning), Fodor, 1998). Referential theories may be accompanied by any number of theories of how reference is (non-descriptively) determined,[9] but the important thing for our purposes is that they have in common a principle of concept individuation.

The referentialist maintains that when concepts are wielded in thought, they serve to identify something (an individual, a kind) rather than to describe it (Millikan, 1999). Successful use of a concept requires mainly that it lock onto, or track, the right thing. An analogy can be made with names in a language, whose purpose is primarily to identify an individual as subject matter. A name can do the job of any number of descriptions, without being equivalent to them:

Douglas: You know that guy?
Adam: Which guy?
Douglas: The one who was in the movie with the silly cartwheels at the end.
Adam: What?
Douglas: You know, the guy with dark hair, always a serious expression, who played one half of a pair of dumbos in his first real movie.
Adam: I still don't know who you mean.
Douglas: He's Canadian, plays the bass in some band, he's a really good actor.[10] And he says "dude" all the time.
Adam: Oh! You mean Keanu Reeves!

Once Douglas has made Adam lock onto the right individual, the conversation can proceed. This could have occurred in any number of ways. Any number of various descriptions could have helped Adam lock onto the right individual. Also, Adam need not have known Reeves' name; it could have just come to him, "Oh, yeah, I know the guy you mean." Or they could have both known his name, and the lock would have been achieved much more quickly. In thought, just as in conversation, sometimes the point of representing something is to obtain a lock on it, in order to make predictions, test an hypothesis, or learn something new about *that very thing* (or that very kind of thing), whatever descriptions it happens to satisfy.

So the referentialist thinks that concepts are individuated by what they lock onto, rather than how that lock is achieved in particular instances. (Thus two people

[8] See Lycan (2000) for an introduction to the issue, though in the (more traditional) context of linguistic meaning.

[9] Some possible reference determining relations include: causation (Kripke, 1972/1980), carrying information (Dretske, 1981/1999), having the function of carrying information (Dretske, 1988), and "asymmetric dependence" (Fodor, 1987).

[10] Note that the descriptions associated with a concept need not all be true in order for the concept to refer.

could have the same concept of dogs, even if they recognize dogs by different means.) SINBAD is naturally interpreted as supporting this kind of concept. SINBAD cells allow an organism to lock onto sources of correlation (like individuals and kinds), but they do so by instantiating many different functions in their dendrites. These functions vary over time, and they need not even be that accurate, especially during learning; nor need the cell be perfectly tuned to the source of correlation that is its referent. So it would make sense to individuate a concept realized by a SINBAD cell by its referent, rather than by an ever changing set of dendritically mediated descriptions. If there is some source of correlation that *explains* the degree to which a cell's dendrites have succeeded in matching their activities (even if they haven't been all that successful), the cell represents that source. Though each cell has a set of correlated 'descriptions' embodied in its dendrites, reference is not *defined* by those descriptions. Rather, the cell refers to whatever is the source or explanatory ground of those (possibly inaccurate) correlated descriptions. Thus a referentialist account of concepts fits well with the SINBAD mechanism.

The third type of theory of concepts that coheres with an aspect of the SINBAD mechanism is the "theory-theory." If the cells in a SINBAD network refer to sources of correlation, the complex associations that exist between cells due to the functions their dendrites instantiate ought to mirror real relations that exist between the sources of correlation that are those cells' referents. This is indeed the case. Our current modeling results show that the effect is enhanced when the cells are connected laterally, so they can make use of one another's outputs (Favorov and Ryder, in preparation). Naturally, interactions between sources of correlation follow regular patterns, indeed they form a large class of those regularities SINBAD cells are apt to discover. Lateral connections allow the network to reflect these regularities more easily, so that the complex relations amongst sources of correlation in the environment are mirrored in the complex associative structure of the network. When the sources of correlation related by a set of interrelated cells form a distinguishable domain of inquiry, those cells may be thought of as embodying a *theory* of that domain. The 'theory-theory' of concepts (Carey, 1985; Murphy and Medin, 1985), an example of an inferential role theory of representational content (Block, 1986; Peacocke, 1992), says that the content of a representation is determined by the inferential role that it plays in a theory. SINBAD cells play inferential roles, by virtue of the dendritically mediated relations they enter into with other cells, both pre- and post-synaptic relative to it. Thus, just as for the descriptionist and referential views of concepts, SINBAD networks exhibit properties central to the theory-theory.

The complex functions embodied in a SINBAD cell's dendrites may be thought of as descriptions eventually linking a concept with the sensory and motor periphery, or as mediating inferences that locate a concept within a theory of a particular domain. By way of its dendritic functions, a cell locks (perhaps imperfectly) onto a source of correlation, an individual or kind the cell attempts to track.

This source of correlation can be used to individuate the concept a cell realizes, and is appropriately thought of as its referent. We tentatively propose that the descriptionist, referentialist, and inferential role theories of concepts draw attention to different aspects of SINBAD, which is the single underlying mechanism of conceptualization. That is, we take these varying theories of concepts to be evidence for SINBAD, *via* an inference to a unifying explanation.

7. Association and Reasoning in SINBAD Networks

The possibility of applying inferential role theories to SINBAD networks brings us to the other problem that Hume's associationism faced. The problem we have already addressed is the naturalization of abstraction: Hume's theory of abstraction was not naturalistic, whereas representation acquisition in SINBAD certainly is. The remaining problem is the naturalization of reason. In this case, the problem was not that Hume's theory failed to be naturalistic, rather the mechanism he proposed – simple pairwise association – is not powerful enough to explain reasoning. We believe that SINBAD networks give us a much more promising mechanism to serve as the reductive base in a new associationist explanation of the capacity to reason.

The problem with simple pairwise association is that although it can explain some transitions in thought, it can explain only a limited number of *rational* transitions. Hume's theory is a mechanization of inductive reasoning, but rational progression from thought to thought is not always of the inductive type. Not only can we reason from "most of the oranges I've seen have been orange", to "this one [seen in the dark] is likely orange"; we can reason from "this fruit is either an orange or an apple" and "it's not an apple" to "it's an orange." This instance of deductive reasoning cannot be reconstructed using simple pairwise association, since pairwise association cannot *exclude* properties from a set, it can only include them. Adding inhibition to pairwise association will go some way to implementing negation, but implementing more complex reasoning based on AND, OR, and XOR require a more complex solution, to say nothing of reasoning with non-linear continuous functions (e.g. being of extreme temperature).

Fodor (1983) conceives of a dressed-up, modern, learning-theoretic, computational associationist. This kind of associationist does not limit himself to simple pairwise association. He postulates other associative relations: e.g., the logical functions *OR*, *XOR*, and more complex functions, making for a small set of associative relations rather than only one. This gives the associationist a fighting chance to explain reasoning, i.e. rational transitions in thought. Fodor's main complaint about computational associationism is as follows. The associationist is forced to admit more complex operations as fundamental in order to account for more complex mental structure. But an acknowledgment of the complexity of mental structure conflicts with the associationist's account of ontogeny (pp. 32–34):

> In short, as the operative notion of mental structure gets richer, it becomes increasingly difficult to imagine identifying the ontogeny of such structures

> with the registration of environmental regularities .... To put the point in a nutshell, the crucial difference between classical and computational associationism is simply that the latter is utterly lacking in any learning theory.

That is, he thinks the classical associationists had a learning theory (association by the constant conjunction of raw stimuli), but that it could not account for rational transitions in thought. 'Computational' associationism, by contrast, has a better chance of accounting for reason, but it is mysterious how a system could *acquire* this mental structure by extracting regularities from the environment.

SINBAD removes the mystery by showing how complex functions and more abstract regularities can be extracted from the environment by an associationist mechanism. Not only can dendrites instantiate complex functions, but the SINBAD learning mechanism shows how these complex functions may be extracted from nature's order. An internal model of the environment, as implemented by a SINBAD network, mirrors the complex relations that exist amongst sources of correlation. In both networks based on simple pairwise correlation and SINBAD networks, representations of states of affairs involving parts of the network will tend ramify throughout the entire network; this is how the associationist proposes to implement reasoning. For example, in the network that learned about sinks (mentioned in section 5), when deprived of information about the temperature of the water coming from the tap, the network reports what temperature it should be, given the state of the other variables (Favorov and Ryder, in preparation). (It reports this *via* the cell in the network that tuned to temperature.) The difference between simple pairwise associationism and SINBAD is that in a SINBAD network, these ramifications can implement complex reasoning (e.g. multivariate Boolean logic and inference to the best explanation[11]), whereas in a simple pairwise associationist network, they can implement only simple induction. It remains to be seen just what types of reasoning can be plausibly realized by the particular version of SINBAD found in the brain; this will depend, in particular, on the precise nature of the cortical synaptic rule (or rules).

8. Conclusion

In this paper we have proposed that pyramidal cells of the cerebral cortex are associative/computational devices designed to discover orderly, predictable relations in their inputs. We interpret the activity of SINBAD networks in the following

[11] The learning operation performed by SINBAD cells may be thought of as a basic form of inference to the best explanation, or 'abduction': the cell 'postulates' a source for the correlations discovered by its dendrites, just as Mendel postulated the existence of genes to explain the correlations that he observed. After representation acquisition, activity ramifying through the network will implement a similar sort of operation, where inputs about surface variables are given a "deep" interpretation as products of the operation of hidden variables. Thus it may be appropriate to think of SINBAD as an entirely new sort of naturalistic theory: while Hume's theory mechanized induction, and Turing's mechanized deduction, SINBAD mechanizes *abduction*.

manner. A pyramidal cell's dendrites discover correlated functions of environmental variables, and the cell thereby tunes to the source of those correlations. Thus a new representation is created, in a way that solves the problem with Hume's abstraction by ensuring that newly created representations are predictively related to others. Such new representations serve as inputs to cells elsewhere in the network, allowing them to discover more deeply hidden sources of correlation. The network as a whole develops into an internal model of the environment. Since the dendrites of cortical pyramidal cells can implement nonlinear functions, this internal model can be much more complex than has previously been accounted for naturalistically. Activity that begins at the periphery (one's sensory evidence) will ramify through the network in a way that implements not only simple induction, but deductive reasoning and inference to the best explanation as well. This shows how it is possible for a simple, biologically realistic mechanism to learn the complexities of nature's order, and use this knowledge for the vitally important task of prediction.

Acknowledgements

This work was supported in part by ONR grant N00014-95-1-0113 and NIH grant NS35222. DR was supported by the Social Sciences and Humanities Research Council of Canada.

References

Abeles, M., 1991: *Corticonics*, Cambridge University Press, Cambridge.

Barlow, H.B., 1992: The biological role of neocortex, in A. Aertsen and V. Braitenberg (eds), *Information Processing in the Cortex*, Springer, Berlin, pp. 53–80.

Becker, S., 1995: JPMAX: Learning to recognize moving objects as a model-fitting problem, *Advances in Neural Information Processing Systems* **7**, 933–940.

Becker, S., 1996: Mutual information maximization: Models of cortical self-organization, *Network: Computation in Neural Systems* **7**, 7–31.

Becker, S., 1999: Implicit learning in 3D object recognition: The importance of temporal context, *Neural Comp.* **11**, 347–374.

Becker, S. and Hinton, G.E., 1992: A self-organizing neural network that discovers surfaces in random-dot stereograms, *Nature* **355**, 161–163.

Bienenstock, E.L., Cooper, L.N. and Munro, P.W., 1982: Theory for the development of neuron selectivity: Orientation specificity and binocular interaction in visual cortex, *J. Neurosci.* **2**, 32–48.

Block, N., 1986: Advertisement for a semantics for psychology, *Midwest Studies in Philosophy* **10**, 615–678.

Braitenberg, V., 1978: Cortical architectonics: General and areal, in M.A.B. Brazier and H. Petsch (eds), *Architectonics of the Cerebral Cortex*, Raven, Philadelphia.

Brooks, R., 1991: Intelligence without representation, *Artif. Intell.* **47**, 139–159.

Burnod, Y., 1988: *An Adaptive Neural Network: The Cerebral Cortex*, Masson, Paris.

Carey, S., 1985: *Conceptual Change in Childhood*, MIT Press, Cambridge, MA.

Cauller, L., 1995: Layer I of primary sensory neocortex: Where top-down converges upon bottom-up, *Behav. Brain Res.* **71**, 163–170.

Chalmers, D., 1996: *The Conscious Mind: In Search of a Fundamental Theory*, Oxford University Press, Oxford.

Clark, A. and Thornton, C., 1997: Trading places: Computation, representation, and the limits of uninformed learning, *Behav. Brain Sci.* **20**, 57–90.

Dennett, D.C., 1994: The practical requirements for making a conscious robot, *Philosophical Transactions of the Royal Society* **A349**, 133–146.

Deuchars, J., West, D.C. and Thomson, A.M., 1994: Relationships between morphology and physiology of pyramid-pyramid single axon connections in rat neocortex in vitro, *J. Physiol. (Lond.)* **478**, 423–435.

Dretske, F., 1981/1999: *Knowledge and the Flow of Information*, CSLI Publications, Stanford.

Dretske, F., 1988: *Explaining Behavior*, MIT Press, Cambridge, MA.

Edelman, G.M., 1987: *Neural Darwinism: The Theory of Neuronal Group Selection*, Basic Books, New York.

Favorov, O.V. and Kelly, D.G., 1996: Local receptive field diversity within cortical neuronal populations, in O. Franzen, R. Johansson and L. Terenius (eds), *Somesthesis and the Neurobiology of the Somatosensory Cortex*, Birkhauser, Basel, pp. 395–408.

Favorov, O.V., Ryder, D., Hester, J.T., Kelly, D.G. and Tommerdahl, M., 2001: The cortical pyramidal cell as a set of interacting error backpropagating dendrites: A mechanism for discovering nature's order, in R. Hecht-Nielsen and T. McKenna (eds), *Theories of the Cerebral Cortex*, Springer-Verlag, Berlin, in press.

Feldman, M.L., 1984: Morphology of the neocortical pyramidal neuron, in A. Peters and E.G. Jones (eds), *Cerebral Cortex*, vol. 1, Plenum Press, New York, pp. 123–200.

Fodor, J.A., 1983: *Modularity of Mind*, MIT Press, Cambridge, MA.

Fodor, J., 1987: *Psychosemantics*, MIT Press, Cambridge, MA.

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

Fodor, J.A. and Pylysyn, Z., 1988: Connectionism and cognitive architecture: A critical analysis, *Cognition* **28**, 3–72.

Gawne, T.J., Kjaer, T.W., Hertz, J.A. and Richmond, B.J., 1996: Adjacent visual cortical complex cells share about 20% of their stimulus-related information, *Cereb. Cortex* **6**, 482–489.

Gelman, S.A. and Coley, J.D., 1991: Language and categorization: The acquisition of natural kind terms, in S.A. Gelman and J.P. Byrnes (eds), *Perspectives on Language and Thought*, Cambridge University Press, Cambridge.

Grossberg, S., 1974: Classical and instrumental learning by neural networks, *Progress in Theoretical Biology* **3**, 51–141.

Grossberg, S., 2000: The complementary brain: Unifying brain dynamics and modularity, *Trends in Cognitive Sciences* **4**, 233–246.

Hancock, P.J.B., Smith, L.S. and Phillips, W.A., 1991: A biologically supported error-correcting learning rule, *Neural Comp.* **3**, 201–212.

Hartley, D., 1749/1970: *Observations on Man*, selections in Robert Brown (ed.), *Between Hume and Mill: An Anthology of British Philosophy 1749–1843*, Random House, New York.

Hume, D., 1740/1978: *A Treatise of Human Nature*, in L.A. Selby-Bigge (ed.), Oxford University Press, Oxford.

Jackson, F., 1998: *From Metaphysics to Ethics: A Defence of Conceptual Analysis*, Oxford University Press, Oxford.

Johnston, D., Hoffman, D.A., Colbert, C.M. and Magee, J.C., 1999: Regulation of back-propagating action potentials in hippocampal neurons, *Curr. Opin. Neurobiol.* **9**, 288–292.

Kant, I., 1787/1996: *Critique of Pure Reason*, Hackett, Indianapolis.

Katz, J., 1972: *Semantic Theory*, Harper & Row, New York.

Kripke, S., 1972/1980: *Naming and Necessity*, Blackwell, Oxford.

Lettvin, J., 1988: *Forward to W.S. McCulloch*, *Embodiments of Mind*, MIT Press, Cambridge, Mass.

Lycan, W.G., 2000: *Philosophy of Language*, Routledge, London.

Magee, J.C. and Johnston, D., 1997: A synaptically controlled, associative signal for Hebbian plasticity in hippocampal neurons, *Science* **275**, 209–213.

Malach, R., 1994: Cortical columns as devices for maximizing neuronal diversity, *TINS* **17**, 101–104.

Malinow, R., Mainen, Z.F. and Hayashi, Y., 2000: LTP mechanisms: From silence to four-lane traffic, *Curr. Opin. Neurobiol.* **10**, 352–357.

Markram, H., Lubke, J., Frotscher, M., Roth, A. and Sakmann, B., 1997a: Physiology and anatomy of synaptic connections between thick tufted pyramidal neurones in the developing rat neocortex, *J. Physiol. (Lond.)* **500**, 409–440.

Markram, H., Lubke, J., Frotscher, M. and Sakmann, B., 1997b: Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs, *Science* **275**, 213–215.

Markram, H., Pikus, D., Gupta, A. and Tsodyks, M., 1998: Potential for multiple mechanisms, phenomena and algorithms for synaptic plasticity at single synapses. *Neuropharmacology* **37**, 489–500.

McGuire, B., Gilbert, C.D., Wiesel, T.N. and Rivlin, P.K., 1991: Targets of horizontal connections in macaque primary visual cortex, *J. Comp. Neurol.* **305**, 370–392.

Mel, B.W., 1994: Information processing in dendritic trees, *Neural Comp.* **6**, 1031–1085.

Millikan, R., 1999: A common structure for concepts of individuals, stuffs, and real kinds: More mama, more milk, and more mouse, in in E. Margolis and S. Laurence (eds), *Concepts: Core Readings*, MIT Press, Cambridge, MA, pp. 525–547.

Minsky, M. and Papert, S., 1988: *Perceptrons*, 3rd edition, MIT Press, Cambridge, MA.

Mountcastle, V.B., 1978: An organizing principle for cerebral function, in G.M. Edelman and V.B. Mountcastle (eds), *The Mindful Brain*, MIT Press, Cambridge, MA, pp. 7–50.

Murphy, G., and Medin, D., 1985: The role of theories in conceptual coherence, *Psych. Rev.* **92**, 289–316.

Paulsen, O. and Sejnowski, T.J., 2000: Natural patterns of activity and long-term synaptic plasticity, *Curr. Opinion Neurobiol.* **10**, 172–179.

Peacocke, C., 1992: *A Study of Concepts*, MIT Press, Cambridge, MA.

Phillips, W.A. and Singer, W., 1997: In search of common foundations for cortical computation, *Behav. Brain Sci.* **20**, 657–722.

Pinker, S., 1997: *How the Mind Works*, W.H. Norton, New York.

Putnam, H., 1975: The Meaning of 'Meaning', in K. Gunderson (ed.), *Language, Mind and Knowledge*, Minnesota Studies in Philosophy of Science vol. 7, University of Minnesota, Minneapolis.

Quartz, S.R. and Sejnowski, T.J., 1997: The neural basis of cognitive development: A constructivist manifesto, *Behav. Brain Sci.* **20**, 537–596.

Quine, W., 1953: Two Dogmas of Empiricism, in *From a Logical Point of View*, Harvard University Press, Cambridge, MA, pp. 20–46.

Rey, G., 1997: *Contemporary Philosophy of Mind*, Blackwell, Oxford.

Rosch, E. and Mervis, C., 1975: Family Resemlances: Studies in the internal structure of categories, *Cogn. Psych.* **7**, 573–605.

Rosenberg, J.R., 1997: Connectionism and cognition, in J. Haugeland (ed.), *Mind Design II*, MIT Press, Cambridge, MA, pp. 293–308.

Rumelhart, D.E., Hinton, G.E. and Williams, R.J., 1986: Learning internal representations by error propagation, in D.E. Rumelhart, J.L. McClelland and PDP Research Group (eds), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, MIT Press, Cambridge, MA, vol. 1, pp. 318–362.

Russell, B., 1914/1993: *Our Knowledge of the External World*, Routledge, London.

Schuz, A., 1992: Randomness and constraints in the cortical neuropil, in A. Aertsen and V. Braitenberg (eds), *Information Processing in the Cortex*, Springer, Berlin, pp. 3–21.

Segev, I., Fleshman, J.W. and Burke, R.E., 1989: Compartmental models of complex neurons, in C. Koch and I. Segev (eds), *Methods in Neuronal Modeling*, MIT Press, Cambridge, MA, pp. 63–96.

Sejnowski, T.J., 1977: Storing covariance with nonlinearly interacting neurons, *J. Math. Biol.* **4**, 303–321.

Singer, W., 1995: Development and plasticity of cortical processing architectures, *Science* **270**, 758–764.

Smith, E. and Medin, D., 1981: *Categories and Concepts*, Harvard University Press, Cambridge, MA.

Spruston, N., Schiller, Y., Stuart, G. and Sakmann, B., 1995: Activity-dependent action potential invasion and calcium influx into hippocampal CA1 dendrites, *Science* **268**, 297–300.

Stroud, B., 1977: *Hume*, Routledge, London.

Stuart, G., Spruston, N., Sakmann, B. and Hausser, M., 1997: Action potential initiation and backpropagation in neurons of the mammalian CNS, *TINS* **20**, 125–131.

Svoboda, K., Denk, W., Kleinfeld, D. and Tank, D., 1997: *In vivo* dendritic calcium dynamics in neocortical pyramidal neurons, *Nature* **385**, 161–163.

Thomson, A.M. and Deuchars, J., 1994: Temporal and spatial properties of local circuits in neocortex, *TINS* **17**, 119–126.

Willshaw, D.J. and Dayan, P., 1990: Optimal plasticity from matrix memories: What goes up must come down, *Neural Comp.* **2**, 85–93.