Noise and the Emergence of Rules in Category Learning: A Connectionist Model

Rosemary A. Cowell[1] & Robert M. French[2]

[1]Department of Psychology, UCSD, 9500 Gilman Drive #0109, La Jolla, CA 92093-0109, USA
[2]LEAD-CNRS (UMR5022), Pôle AAFE, Esplanade Erasme, Université de Bourgogne, 21000 Dijon, France

Corresponding authors:
Rosemary Cowell: rcowell@ucsd.edu
Robert French: robert.french@u-bourgogne.fr

We present a neural network model of category learning that addresses the question of how rules for category membership are acquired. The architecture of the model comprises a set of *statistical learning* synapses and a set of *rule-learning* synapses, whose weights, crucially, emerge from the statistical network. The network is implemented with a neurobiologically plausible Hebbian learning mechanism. The statistical weights form category representations on the basis of perceptual similarity, whereas the rule weights gradually extract rules from the information contained in the statistical weights. These rules are weightings of individual features; weights are stronger for features that convey more information about category membership. The most significant contribution of this model is that it relies on a novel mechanism involving feeding noise through the system to generate these rules. We demonstrate that the model predicts a cognitive advantage in classifying perceptually ambiguous stimuli over a system that relies only on perceptual similarity. In addition, we simulate reaction times from an experiment by Thibaut et al. (1998), in which both perceptual (i.e., statistical) and rule based information are available for the classification of perceptual stimuli.

Index Terms: categorization, rule emergence, noise, neural network

I. INTRODUCTION

The categorization of objects on the basis of their visual attributes is a cognitive capacity fundamental to our survival. Human adults, as well as infants over one year of age are able to categorize objects based not only on the statistical structure of categories of observed objects, but also by making use of rules derived from that structure. Rules have the intrinsic advantage of radically reducing cognitive load: if an object can be categorized by paying attention to only one or two of its features, instead of a great many, cognitive resources can be freed up for other tasks. The ontological status of rules in a connectionist modeling framework has from the outset been a hotly debated topic (see, for example, Seidenberg & McClelland, 1989; Pinker & Prince, 1988; Smolensky, 1988; Marcus et al., 1999). In this paper we have chosen a conciliatory point of view — namely, that rules do have a distinct ontological status compared to purely statistical learning mechanisms, but that these rules, in general, emerge from the statistical learning substrate.

Our usage of the expression *rule for categorization* is not the commonly used 'necessary and sufficient condition' for categorization, but, rather, it is a condition that is generally sufficient for category membership. In acquiring knowledge of such a 'quasi-sufficient' condition, the observer must go from attending to all perceptual features to attending to only a small subset of features that appear uniquely in a given category: 'category-diagnostic' features. To do this, one must learn which features to attend to and which to ignore. In other words, during the acquisition of the rule, features associated with several categories must drop out of the category representation. Rules of this nature might include: animals with beaks are birds; animals with gills are fish; animals with opposable thumbs are primates; and so on. And even though, for example, opossums, koalas and giant pandas also have opposable thumbs, these rules are generally true. Not only this, such rules are adaptive because they free up cognitive resources and, most importantly, they can be extracted from the feature statistics of primates and birds.

Finally, and critically, once a rule has been acquired, an agent should not rely solely upon this knowledge. Rather, rules emerge from knowledge about the distribution of perceptual features in the categories experienced and, once acquired, are used in tandem with the original perceptual knowledge. The two systems may complement or compete with each other, depending on the category structure and the objects encountered. For most stimuli, the rule and perceptual knowledge will be in agreement and will reinforce each other. But for some stimuli they will compete, producing conflict and yielding slower responding.

We present a neural network model with two sets of synapses: a statistical learning set operating in tandem with a rule learning set that extract rules from the statistical set. (A preliminary version of this model is described in Cowell & French, 2007.) These two sets of synapses connect the same input and output nodes, therefore their effects interact extremely closely. An important and unique novel contribution of this model is that the gradual emergence of rules is due simply to the presence of noise in the system. It has long been known that “rather than [being] merely a nuisance, noise in biological systems is a useful property” (Traynelis & Jaramillo, 1998). Noise has been proposed as a mechanism for long-term memory consolidation (Abraham & Robins, 2005; Ans & Rousset, 2000; French, 1997) and has recently been shown to play a key role in optimizing population coding in the brain (Ma et al., 2006; Averbeck et al., 2006). In the present paper we suggest that it may also play a key role in “implicit” rule learning (i.e., rule learning where one is not explicitly told the rule.)

II. CONTRIBUTIONS OF THE MODEL

Arguably, the most successful neurobiologically grounded model of category learning in the literature comes from the work of Ashby and colleagues (Ashby et al., 1998; Ashby & Ell, 2001). This connectionist model is based upon the idea of competition between verbal and implicit

systems (COVIS) and advocates multiple systems for category learning. Our model shares certain properties with COVIS, for example that it possesses multiple systems (statistical weights and rule weights), the outputs of which are combined in order to produce the overall categorization response. The COVIS model of multiple systems for category learning is arguably a more complete and more comprehensively-tested model of categorization behavior than the work we present here. However, in COVIS, the rule learning process consists in selecting between alternative verbalisable rules, with no mechanism offered as to how the candidate rules are generated from existing knowledge in the first instance. A key contribution of our system is that it offers an explanation of how the simple, verbalisable rules that are assumed in a system such as COVIS could be brought into existence. That is, rather than using "off-the shelf" rules, our model extracts simple rules from its perceptual system. The development of rules in the system is, we believe, the principal contribution of our model to the categorization literature.

One consequence of the rule development mechanism in our model is that the relative contributions of the rule and implicit systems differ from those assumed in COVIS. While Ashby and colleagues argue that the verbal (rule based) system dominates category learning initially and categorization behavior may be dominated by implicit knowledge only later in learning, our system necessitates that statistical (implicit) knowledge comes online first, before any rules may be extracted and come to govern responding.

The two systems in our model work in concert as much as in competition. Knowledge in the two sets of weights develops in parallel; indeed, one system depends upon the other and, although the two routes can produce conflicting responses, they share common mechanisms. The interdependence of our two systems contrasts with COVIS, and with other dual-route models of cognitive function in which the two systems proposed are independent (e.g., Coltheart et al., 1993).

Figure 1. A kiwi, as pictured here, is likely to cause conflict in a categorization task because of the incongruity between its beak, which triggers the rule: "if *beak*, then *BIRD*", and its visual similarity to a rodent.

In addition, our model provides predictions of reaction time for categorization. The reaction time to a particular stimulus depends, in part, on the degree of conflict between the responses from the perceptual and rule based systems. Often, the appearance of a category-determining feature (e.g., *beak*) is highly correlated with other features that may not be category-determining (e.g., *two legs*, *wings*, *small size*, etc.). For most instances of an animal belonging to the category *BIRD*, the response of the rule system (which infers *BIRD* from the presence of *beak*) will agree with the response of the perceptual system (which infers *BIRD* from the presence of *beak*, *two legs*, *wings*, *small size*, *feathers*, etc.). However, an instance such as a kiwi (Fig. 1) causes conflict in the system, because although the *beak* signals *BIRD* to the rule system, the furry-looking feathers, lack of visible wings, small size and general rodent-like appearance signal *VOLE* to the perceptual system. We would be expected to take longer than normal to classify the kiwi as a *BIRD*, in spite of its category-defining feature *beak*.

III. OVERVIEW OF THE MODEL

The network possesses two sets of units – inputs and outputs – and *two* sets of weights between the two sets of units: a 'statistical' weight set and a 'rule' weight set, which are distinguished by the type of Hebbian update they receive. The two weights sets have identical connectivity, and could be thought of as two types of synapse, or indeed simply as two types of learning occurring at the same synapse. Activation in the output units is determined by passing inputs through both sets of weights (which sums the effects of the two types of synapse, or two types of learning).

The 'statistical learning' weights learn the distributions of perceptual attributes of the stimuli in each category, whereas the 'rule' weights derive their rules by learning from the transformation of noise activation at the input units by the statistical weights. The model outputs both a 'category response' and a 'reaction time', as described later.

The architecture and learning principles of the network are based upon a Kohonen network (Kohonen, 1993) in that output units are connected by lateral weights that are locally excitatory and distally inhibitory. The statistical synapses can be learned by self-organization, or may be subject to intermittent or consistent feedback; all learning of both statistical and rule synapses is Hebbian. The rule learning synapses exploit noise in the input layer, which may travel via both sets of synapses to the outputs but engenders learning only in the rule weights; in this way, the rule synapses determine which input features are sufficient for determining category membership. The network is implemented in a neurobiologically plausible manner, similar to Kohonen (1993), using leaky-integrator neurons. It has been suggested that processing of this type occurs in visual cortex (Kohonen, 1993; Spitzer, 1999; O'Reilly & Munakata, 2000).

The rule extraction mechanism of our model is based on the following principle. If a particular input (i.e., feature) unit in the statistical learning network has a strong weight connecting it to only one category output node and weak weights to all other category output nodes, this means that feature will activate one, and only one, category node. The presence of this feature tells us that the stimulus item must belong to the category whose output node it activates; we refer to such features as 'diagnostic' features.

For example, in Fig. 2 the weight in the statistical learning network between *beak* and *BIRD* will become large during training because all birds have beaks. Every time a beak is encountered, the category membership of the item with a beak will be *BIRD*. So the *beak*-*BIRD* link will become very strong, while the weights between *beak* and

any other category node will remain small. *Beak* is a diagnostic (i.e., category-determining) feature. On the other hand, the feature *eyes* is shared by birds, bats, and cats; so for *eyes*, no link from that feature unit to any one category node will be strong in comparison with links from it to all other categories. *Eyes* are a category-irrelevant feature. The essence of our rule network is that it detects when an individual feature possesses a strong link to only one category, allowing the network to conclude that that feature must constitute a rule for categorization.

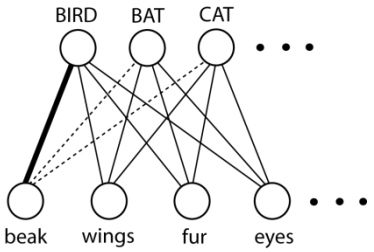

Figure 2. Any animal for which the first feature (*beak*) is active is a *BIRD*. In the statistical learning network, the weight between *beak* and *BIRD* is large, while the *beak-BAT*, *beak-CAT*, etc. weights are small. This is the information that is extracted by the rule network.

As we will explain in detail below, the model operates by exploiting noise in the system to allow transformations of activation by the statistical synapses to be read out and copied, selectively, by the rule synapses. The category response of the network to a novel stimulus is determined at the output units, which are activated via both the statistical learning and the rule learning weights.

IV. MODEL ARCHITECTURE

The general architecture of the model is shown in Fig. 3. The model consists of a set of 'perceptual feature' nodes in the input layer, which are connected to a set of category nodes in the output layer, via two sets of weighted feedforward connections. The output layer is a one-dimensional array of processing units that receives inputs from stimuli or noise and implements lateral excitation and inhibition between neighboring units (Fig. 4). The statistical weights and the rule weights that connect input (feature) units to output (category) units are both incrementally adapted via a Hebb-type learning rule. The output activations are strongly influenced by the lateral connectivity, such that only one region of the output layer remains active once activations have settled, following lateral inhibition.

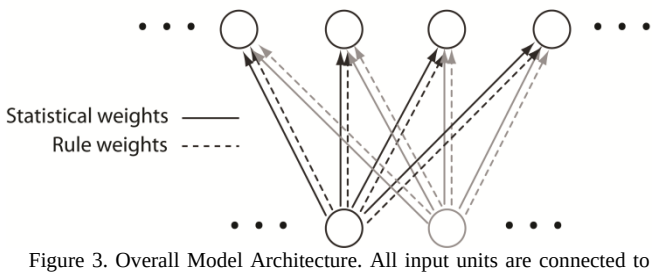

Figure 3. Overall Model Architecture. All input units are connected to all output units via two feedforward sets of weights: statistical and rule weights (solid and dashed lines, respectively). For simplicity, only 2 units are shown in the input layer and 4 in the output layer; in simulations, there were 10 input units and 8 or 9 output units. Weights from the left and right input nodes are shown in black and grey, respectively, for clarity only; there was no difference between these in the model. Lateral connectivity not shown; see Fig. 4.

The neural network model is formulated in a biologically plausible manner. In particular, the lateral inhibition is implemented with neuron-like properties, as illustrated in Fig. 4, which provides extra detail for which there was insufficient space in Fig. 3. Lateral interactions are implemented directly between individual output units. Each unit in the output layer receives excitatory input from the input layer, and inhibitory input from both an interneuron and from other collateral units. The activations of units in the output layer are then calculated dynamically and simultaneously, so that each unit's activation evolves according to the input it receives from other units – via both feedforward and collateral links – some of whose activations are simultaneously being adjusted. Both output units and interneurons are subject to activation leakage. This was implemented as a set of non-linear differential equations similar to those described in Kohonen (1993).

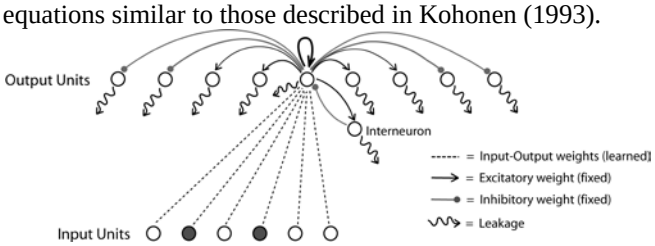

Figure 4. Details of the lateral connectivity on the output units. For clarity, only 6 units are shown in the input layer, whereas there are 10 input units in the model. Also for clarity, only one set of input-output weights is shown from each input node to the central output node; in the model, each input node connects to *all* output nodes with *two* sets of weights (as in Fig. 3). All output units are coupled with an interneuron and have fixed-weight lateral connections to all other output nodes. See Appendix for lateral weight values.

The evolution of output activations in response to input activation differs slightly, depending on whether that activation comes from stimuli or from noise. It is presumed that the lateral interaction processes are influenced by the fact that stimuli are presented for a minimum of several hundred milliseconds, whereas noise activation appears very transiently. Accordingly, when stimuli are presented to the network, we assume sufficient time for the activations of output units and interneurons to interact (via both lateral weights and interneuron-output weights), undergo leakage and co-evolve to a stable state. This process is simulated by finding the numerical solution to a pair of differential equations defining the activation of output units and inhibitory interneurons (see Appendix). In contrast, when noise activation appears on the input layer, it appears as a transient burst, which is not sustained for long enough to allow full evolution. Instead, activation is simply passed to the output units via the feedforward weights, whereupon output activations cycle briefly through the lateral connections before reaching their final values.

V. STATISTICAL LEARNING MECHANISM

For each presentation of a stimulus, the input pattern is clamped to the statistical input layer of the network, resulting in the evolution of a stable and sustained activation pattern across the outputs. It is assumed that Hebbian learning on the statistical synapses requires sustained and simultaneous activation of both the sending and receiving nodes. Therefore these synapses are updated when stimuli are presented, but not when noise activation passes through the network (see Section VI, below, for activation due to noise).

Since stimuli within a category typically share many perceptual features, they come to elicit activation on the

same region of the output layer. Thus, they share a representation and are classified by the statistical learning weights as belonging to the same category.

VI. RULE LEARNING MECHANISM

We assume the occurrence of activation due to noise in the input units, which echoes the activity elicited by stimuli. This noise occurs spontaneously at the input layer, and is interleaved with the presentation of training stimuli. The rule weights in the network exploit this noise activation to extract the 'category-diagnostic' features of stimuli as simple rules. Once learned, the rule synapses map input stimuli onto category representations using only category-determining features. While other algorithms have been developed (e.g., Rossi, 1996; Thomas, van Hulle, & Vogels, 2000) for determining the relative importance of the weights in a Kohonen network, an aim of our model was to implement the extraction of rules with structures and mechanisms that could conceivably arise in the cortex.

The rule synapses feed forward from the input to the output units with exactly the same connectivity as the statistical learning synapses (see Fig. 3). The rule synapses, like the statistical learning synapses, are adjusted during training with a Hebb-type learning rule. However, in the case of the rule weights, it is assumed that learning on these synapses requires asynchronous activation of the sending and receiving nodes. That is, for learning on the rule synapses, the input nodes must be activated first and activation in the output nodes must emerge only after (but soon after) input activation has dissipated.

This is the pattern of activation that occurs when noise activates the input units. Noise activation appears transiently (for a single timestep) on the inputs and is passed to the output units via any synapse (rule or statistical) possessing some connection strength. The activation that results in the output units emerges only after a brief period of cycling through the lateral weights, to resolve the lateral competition; however, by this time, noise activation at the input has dissipated. These conditions being correct for rule-weight learning, there follows an update to any connections between inputs and outputs that were activated, in succession, by the noise activation.

VII. THE EXTRACTION OF RULES USING NOISE

The critical mechanism that allows the rule network to learn diagnostic features for category membership is the way in which noise-triggered activation is propagated through the statistical learning category layer (Figs. 5 and 6). We know of no other model of rule-learning that posits the involvement of noise as a means of rule extraction from a statistical substrate.

We assume that noise generally triggers activation of only one feature (one unit in the input layer) at a time. This assumption is discussed in detail below. The case where noise activates a category-determining feature – in the example above, *beak*, which determines membership of the category *BIRD* – is illustrated schematically in Fig. 5.

First, spontaneous activation is triggered by noise in the input units (1). This activation is passed to the output units via the statistical learning weights (and via the rule weights, once they are learned, but in the example the rule weights are not yet learned) (2). At the outputs, activation cycles very briefly through the lateral weights to resolve lateral competition (3). Note that, by this time, the original noise activation has dissipated at the input units. Finally, Hebbian learning occurs at all connections between active output units and input units that were recently active (4). This results in an increase in the strength of the 'rule weight' from the unit representing the category-determining feature in the input layer to the unit representing the correct category in the output layer. The rule weights are learning the rule for this feature-category pair. Since the statistical weights require simultaneous, sustained co-activation of inputs and outputs, no update of statistical learning weights occurs.

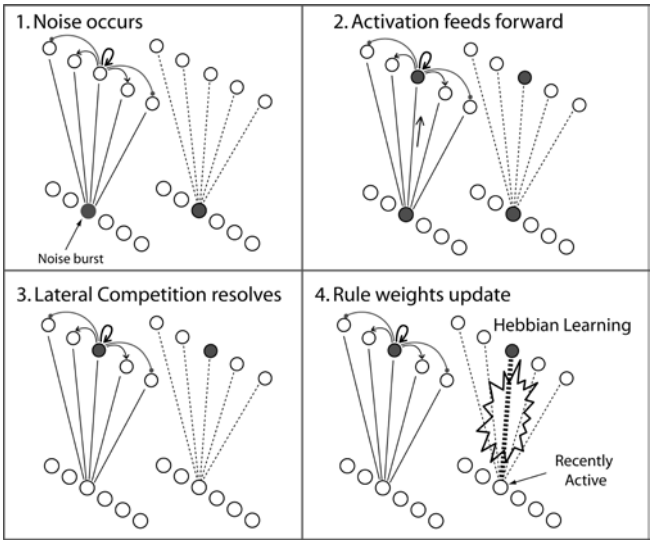

Figure 5. Learning rules for category-diagnostic features by exploiting noise in the system. Input and outputs are depicted *twice* in each panel, in order to illustrate clearly the two weight sets, i.e. the units on the left of each panel are the *same units* as those shown on the right. Statistical weights are shown on the left of each panel as solid lines; rule weights are shown on the right of each panel as dashed lines.

Fig. 6 shows a schematic illustration of the case in which activation triggered by noise occurs in a category-irrelevant feature; in the example given earlier, this might be the feature *eyes*, which does not determine category membership since it is shared by many animals, including *BIRDS*, *BATS*, *DOGS*, and *CATS*. First, spontaneous activation occurs in the input units and activates the feature *eyes* (1). This activation is passed to the output units (2). Next, in (3), lateral competition between output units is resolved by allowing activation to cycle very briefly through the lateral weights. Since the original spread of activation across the output units was broad (because the feature is associated with multiple categories and, thus, several categories are activated in the output units), the effect of the lateral inhibition among output units is to suppress activation everywhere: no-one wins the competition. (Note that this does not occur when a stimulus is presented because the input pattern is clamped to the input layer for a sustained duration, in contrast to the very brief presentation of noise activation). In (4), no activation is present in the output units at the critical time point, shortly after activation at the inputs has dissipated. Therefore, no Hebbian learning occurs in the rule weights on the connection between the input unit corresponding to that feature and the output units corresponding to the categories that are associated with it. In effect, the weights between this feature and the categories possessing the feature have dropped out of the category representations in the rule synapses. The rule weights are left with only those feature-category mappings that are diagnostic for category

membership.

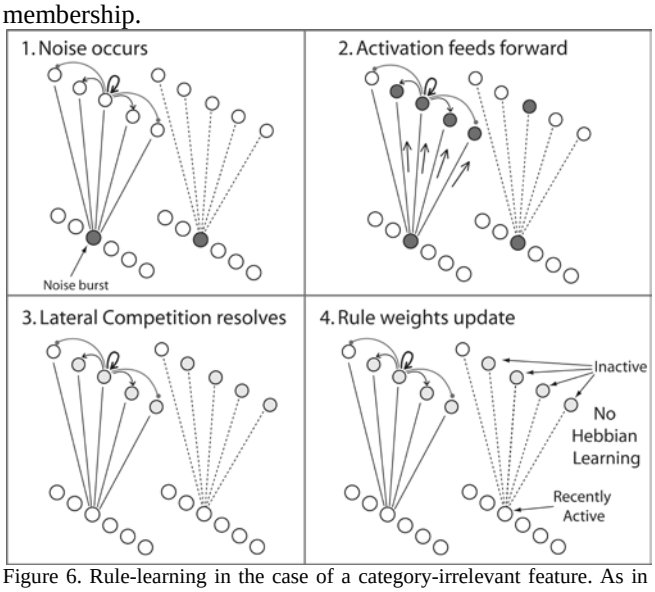

Figure 6. Rule-learning in the case of a category-irrelevant feature. As in Fig. 5, input and output units are depicted twice to show the two weight sets clearly. Statistical weights are solid lines, rule weights are dashed lines.

VIII. NOISE ACTIVATION TO GENERATE RULES

One question that arises from our suggestion that noise drives rule learning is, Why doesn't noise also activate the output nodes of the model? The spontaneous activation driving rule-learning is described as appearing only on the input units, but not elsewhere. We readily admit that noise activation is equally likely to appear in the output units, but we do not model these occurrences because they would not affect the learning processes in the model. Since all connections in the network are feed-forward, noise activation occurring in the output units is never propagated back to the input units. All learning is Hebbian and therefore, since input units are not activated by noise activation triggered in output units, no co-activation dependent learning is induced.

Another question concerning the use of noise as the mechanism for rule learning is, How can we assume that noise will activate only a single input node at a time? In our model, an individual input unit represents a perceptual feature and therefore does not correspond to a single neuron, but rather to a group of neurons whose collective firing stands for a particular perceptual property. What we mean by the occurrence of spontaneous firing in an input unit is that, as a result of spontaneous neural firings in individual neurons, pattern completion is occurring in visual cortex. That is, the group of neurons which together represent a commonly-encountered perceptual feature fire so often together that they form a tightly bound unit, or Hebbian cell assembly (Hebb 1949), in which the simultaneous occurrence of activity in a small subset of the neurons causes the whole representation to be activated. Only a very few neurons in this population need fire for a burst of activity to be generated across the whole group.

We suppose that cell-assemblies that have more recently fired in response to a stimulus are more likely to be subject to spontaneous activity. One can assume that there is residual activation in any recently active cell assembly, so that for a short time after the disappearance of the stimulus, the activation level in the cell assembly remains near its perceptual threshold. Therefore, noise activation in the model occurs only on those features that appear during presentation of the stimuli in the training set.

IX. OUTPUTS OF THE MODEL

After training, the category response behavior of the model is assessed by presenting test stimuli. Since both the statistical- and the rule-learning weights have acquired some connection strength by the end of training, both weight sets contribute in tandem to the activation of the output units. Just as with training stimuli, test stimuli are clamped to the inputs while the output activations evolve according to the coupled differential equations (see Appendix). Through training, each category has become associated with particular output units. Strong activation in these units can be taken as specifying a particular category, but the location of category representations across output units may vary from one trained network to the next. We determine which units correspond to each category by recording the frequency with which each output unit "wins" (attains the maximum activation) in response to each category of stimulus, during the end phase of training. To measure the model's category response on test, we find the identity of the most active output unit and compare it to the frequencies from training: if the winning unit on test matches either of the two most frequently winning units from training, for the category of the test stimulus, the response is 'correct'. (Note that many of the test stimuli are distorted in some way from the category templates seen in training, so that a correct answer is not guaranteed). We examine not only the activations of the output units, but also the number of timesteps it takes for those values to cross a threshold, thus extracting both a 'category response' and a 'reaction time'.

The idea that two systems in categorization operate simultaneously and interact has already been proposed and supported by empirical findings. For example, Keil (1989) has suggested that the categorization behavior he has observed in children and adults arises through, first, the derivation of critical features that can outweigh an object's apparent similarity to members of other categories and, second, the combination of this knowledge of critical features with similarity information. Keil's view is that theoretical relations (i.e., critical features that people use to construct 'theories' of, say, biological entities) dominate over perceptual (i.e., statistical) features. In addition, Allen and Brooks (1991) showed that a rule system may dominate over an perceptual system in categorization, but the latter cannot be suppressed completely when rules are multi-dimensional.

There is a considerable literature examining the time-course of perceptual choice and the mechanisms by which alternative responses compete for control of behavior (see Usher and McClelland, 2001, for a review). We model the production of longer reaction times to stimuli that induce conflict between the statistical and rule weights by examining the time course of evolution of the output unit activations. Because we use a leaky-integrator implementation of the Kohonen algorithm, in which output units compete for the highest activation via lateral inhibition, we are able to use the time it takes for the competition to be resolved – that is, for one unit to cross an activation threshold and 'win' – as a proxy for reaction time. In this way, our reaction time mechanism is reminiscent of the leaky-integrator solution of Usher and McClelland (2001). When more than one unit starts out with a non-

negligible level of activation – that is, when there is conflict as to the correct answer – it takes longer for the competition between units to be resolved.

X. TRAINING AND TESTING THE MODEL

The model is trained by repeated presentation of stimuli belonging to the different categories that it is required to discriminate; there may be two or three such categories. During training, after each presentation of a training stimulus, noise-generated activation is simulated on the input units and this activation is passed through any non-zero feedforward weights to the output units. The activation due to noise is critical to the rule-extraction mechanism. During testing we consider up to three different outputs from the system: (i) the category response of the complete model, (ii) a measure of reaction time of the complete model, and (iii) the category response of a model possessing the statistical learning weights alone, to assess the benefits of possessing the rule-extraction mechanism. Our first simulation serves simply to demonstrate that the system is able to extract rules from a statistical learning substrate and that, for highly ambiguous stimuli, it can provide better categorization performance than a system based on statistical learning alone. Our second simulation demonstrates the ability of the model to simulate reaction time data and a key behavioral phenomenon from a categorization experiment with human subjects (Thibaut et al., 1998) in which both statistical, perceptual information and rules were available.

XI. SIMULATION 1

This simulation shows the rule-extraction mechanism in operation and demonstrates predictions for categorization performance and reaction times. The category structure of the training stimuli contains one-dimensional rules.

	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Cat A	●	○	○	○	○	○	○	○	○	●
Cat B	●	○	○	○	○	●	○	○	○	○
Cat C	○	●	○	○	○	○	●	○	○	○

Table 1. Category structure for Simulation 1. Filled circles represent features that took high values in all instances of stimuli from that category.

A. Training Stimuli

Stimuli were represented as an input vector with ten elements (i.e., features). Each feature may be thought of as some visual-perceptual attribute of an object, represented by a number between 0 (indicating total absence of the feature) and 1 (indicating that the feature is highly salient). Stimuli had two high-valued elements (i.e., between 0.6 and 1) and eight low-valued elements (i.e., between 0 and 0.2). These values differed for each particular category exemplar, but, for example, category C stimuli always had high values on features 2 and 7 and low values elsewhere. The stimuli were divided into three categories, A, B, and C, as shown in Table 1. Categories A and B had an overlapping feature – feature 1. Because feature 1 occurred in both categories A and B it was not diagnostic for either. Each category was defined by at least one sufficient feature: category A by feature 10; category B by feature 6; and category C by features 2 and 7.

B. Training Procedure

Two groups of six networks were initialised: a 'Complete Model' group, which contained both the statistical learning and rule-learning weight sets, and a 'Statistical' group, comprising networks that possessed only the statistical learning weight set. We make the comparison between the complete model and the statistical learning system to demonstrate the effect of adding a rule extraction mechanism on categorization performance.

During training, 400 exemplars from each of the three categories, A, B, and C, were presented in random order to all networks, giving 1200 training stimuli in total. The training protocol involved 7 steps (outlined below). Steps 1-7 were repeated, in order, until all training stimuli had been presented.

1. Present training stimulus
2. Allow activations to evolve dynamically
3. Update statistical weights according to a Hebb-type learning rule requiring simultaneous activation
4. Simulate the presence of noise in one of the previously activated features in the Statistical Input layer
5. Allow activations to evolve dynamically
6. Update rule weights according to a Hebb-type learning rule requiring delayed activation
7. Repeat steps 4 to 6 twice more (3 noise bursts in total)

C. Test Procedure

After training, to investigate whether the acquisition of rules had a demonstrable influence on classification behavior, we tested networks by presenting some "atypical" category exemplars as test stimuli.

D. Test Stimuli

We measured categorization performance and reaction time with four instances each of three templates that defined three novel, atypical stimuli. The templates did not conform exactly to any of the category templates from which training stimuli were generated (see Test Items 1, 2 and 3 in Table 2). The atypical stimuli each contained at least one category-diagnostic feature, but also included another distracter feature that was ambiguous with respect to category membership. Thus, the category membership of Test Items 1, 2 and 3 was determinable from the category-diagnostic feature contained in each, despite the fact that they were atypical examples of their respective categories. Test Item 1 was a member of category C owing to the presence of category-determining feature 7; Test Item 2 was a member of category C owing to the presence of category-determining feature 2; and Test Item 3 was a member of category B owing to the presence of category-determining feature 6. Note that Test Items 1 and 2 had perceptual overlap with categories A and B because of the presence of (non-diagnostic) feature 1. In the same way that Test Items 1 and 2 were distorted example of category C items, Test Item 3 was a perceptually distorted exemplar of a category B item, since it contained a previously unseen feature 3. However, Test Item 3 did not possess features that appear in any other category, i.e. the distortion did not create ambiguity by making the stimulus more perceptually similar to members of another category, as was the case for Test Items 1 and 2.

	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	
Test 1	●	○	○	○	○	○	●	○	○	○	*Cat C*
	●	●	○	○	○	○	○	○	○	○	
	○	○	●	○	○	●	○	○	○	○	

Test 2		*Cat C*
Test 3		*Cat B*

Table 2. Templates for the generation of atypical Test Stimuli 1, 2 and 3 of the simulation. The stimulus features that are 'on' in each test item are shown in dark grey.

For Test Items 1 and 2, we expected networks in the 'Complete Model' group, which possessed rules associating each diagnostic feature with a particular category, to ignore the distracter feature and correctly classify these items as belonging to category C. Conversely, a network with only statistical learning synapses should be misled – at least some of the time – by the distracter feature in Test Items 1 and 2, which renders these stimuli more similar to category A and B items. We expected networks in the 'Statistical' group sometimes to classify Test Items 1 and 2 as members of categories A and B, and sometimes to produce a "hybrid" or weak response not corresponding to any category.

For Test Item 3, an atypical category B exemplar, the distracter feature that distorted the stimulus away from the category B template did not appear during training as part of any stimulus. Therefore, this distracter feature does not render Test Item 3 more perceptually similar to items in other categories and we expected this control test stimulus to produce less misclassification by the statistical learning weights than Test Items 1 and 2.

E. Results

Categorization performance of both network groups on the three novel, atypical test stimuli is shown in Fig. 7. 'Complete Model' networks (possessing both statistical and rule-learning weights) performed well on the categorization of all three atypical test stimuli, classifying them according to the rule rather than being influenced by the perceptual similarity of the test items to other categories. However, networks in the 'Statistical' group performed poorly on Test Items 1 and 2, but well on Test Item 3. A two-way ANOVA (Group x Test Item) was performed on the percent correct scores for the three perceptually distorted stimuli, Test Items 1, 2 and 3. There was a significant effect of Group ($F(1,10) = 39.03$, $p < 0.001$), a significant effect of Test Item ($F(2,20) = 38.55$, $p < 0.0001$), and a significant Test Item x Group interaction ($F(2,20) = 22.96$, $p < 0.0001$).

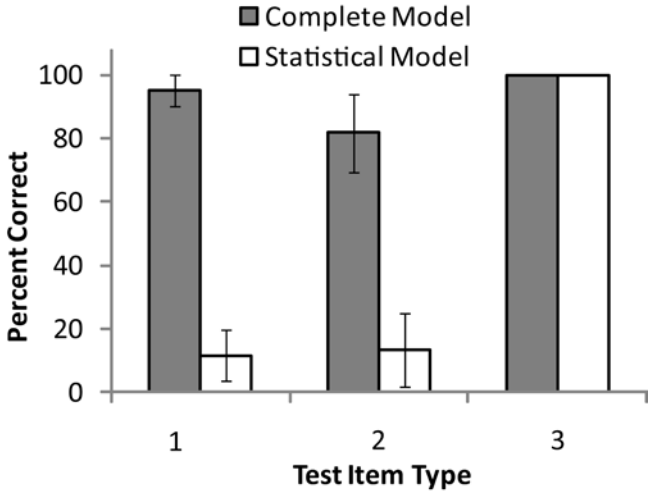

Figure 7. Categorization performance (Percent Correct ± SEM) on Test Items 1, 2 and 3. 'Correct' indicates that the stimulus item was classified in accordance with the rule.

The reaction times of 'Complete Model' networks to the three test stimulus types are shown in Fig. 8. These reaction times are averages over only those trials on which the network produced a correct response. The figure shows that Test Items 1 and 2 – which possessed a non-diagnostic and ambiguous feature – yielded longer reaction times than the 'control' Test Item 3. A repeated measures one-way ANOVA comparing the mean reaction times for Test Items 1, 2 and 3 revealed a significant effect of test stimulus type ($F(2,10) = 40.91$, $p<0.001$). Pairwise group comparisons with Sidak adjustment for multiple comparisons revealed that Item 3 stimuli were classified more quickly than either Item 1 ($P<0.01$) or Item 2 ($p<0.01$) stimuli, but that reaction times for the latter two did not differ ($P=0.996$).

F. Simulation 1 Discussion

Both groups of networks performed perfectly on the "control" test stimulus – Test Item 3, an atypical exemplar of category B. The way in which this stimulus was distorted from the training template for category B left no ambiguity as to which category from training the item most closely resembled; that is, according to either statistical similarity or rules, the item belonged to category B.

Where the two groups differed significantly was on Test Items 1 and 2, for which the 'Statistical' group performed poorly because they were misled by the presence of distracter feature 1. Networks either classified the stimulus as a category A or B item on the basis of perceptual similarity, or were sufficiently misled by the perceptual similarity to categories A and B in the presence of the rule for category C that they were unable to settle upon a representation that corresponded to a category from training. The 'Complete Model' group, which possessed rule weights, did not suffer from this problem because its

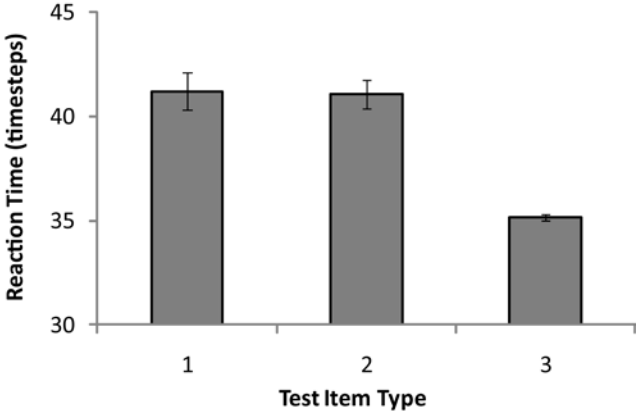

Figure 8. Reaction time ± SEM of 'Complete Model' networks for categorization of Test Items 1, 2 and 3. Only correct responses are included. The unit of reaction time is the number of timesteps in the evolution of output activations until the activation of any node passes threshold.

behavior was guided by the presence of the rules: "if feature 2, then category C" and "if feature 7, then category C". Test Item 3 can be thought of as a "control" test item, because it is distorted from the category training exemplars like Test Items 1 and 2, but there is no ambiguity as to its category identity, because its distracter feature, 3, never appears in training. Critically, there was no difference in categorization performance between the Statistical and Complete Model groups on this stimulus, confirming that the differences seen for Test Items 1 and 2 were not simply due to increased task difficulty. Finally, since there was no ambiguity as to the category membership of Test Item 3 and no conflict between the statistical and rule information for

this item, reaction times for this stimulus were significantly shorter than those for Test Items 1 and 2. This demonstrates that the increased reaction times seen in Fig. 9 for Test Items 1 and 2 is produced by the conflicting information they possess, rather than a non-specific effect of stimulus unfamiliarity.

One further noteworthy result from the simulation is that in the Complete Model, the rule weights have extracted not only a simple "if… then…" rule for classifying the stimuli, but also an OR rule. In the case of Categories A and B, each rule is simple: "if feature 10, then Category A", and, "if feature 6 then Category B". For Category C, however, the model has extracted *two* separate feature rules, which together form a disjunctive (OR) rule of the form: "if feature 2 OR feature 7 then Category C". In categorizing Test Items 1 and 2, the rule network exploits the two halves of this OR rule separately.

XII. SIMULATION 2

Thibaut, Lemaire and Quadri (1998), investigated the learning of a category structure in which both perfectly predictive rule-based information and imperfectly predictive "statistical" information were available for the classification. The study yielded insights into the nature of the interaction between a rule-based and an associative system, such as those implemented in our model. In the study, participants were presented with stimuli from two categories, with feedback as to category membership. Stimuli were abstract, hand-drawn shapes consisting of a horizontal body with four legs protruding downwards from the body. The rule for categorization was contained in the grouping of the legs: either the legs were arranged as two groups of 1 leg and 3 legs (the '1-3' category), or they were as arranged as two groups of 2 legs (the '2-2' category). In addition to this infallible rule, stimuli in the '1-3' category were associated 60% of the time with a rounded upper-body shape. Stimuli in the '2-2' category were associated 60% of the time with an angular, elongated upper-body shape. However, the body shape was not a perfect predictor: 10% of stimuli in each category were associated with the body shape seen 60% of the time in the other category (i.e., 10% of '1-3' stimuli were angular/elongated, and 10% of the '2-2' stimuli were rounded). In addition, 30% of the stimuli in each category were 'neutral', being neither elongated nor rounded. Such a category structure therefore provides perfect rule-based information based upon the leg groupings, in addition to an imperfect "perceptual" information based upon body shape.

Thibaut et al. (1998) also included in their experiment an 'associative phase', in which *only* stimuli with the more commonly appearing body type for that category were presented (i.e. '1-3' rounded stimuli and '2-2' elongated stimuli). This phase was intended to strengthen the association of each body shape type with the category it most often appeared in during the first phase. In fact, this scenario should make the association between body type and category membership stronger, but it may also lead participants to codify body shape as a rule, since during these trials body shape becomes a perfect predictor.

At test, Thibaut et al. (1998) asked participants to classify three types of stimulus: "congruent", "contradictory" and "neutral". Congruent stimuli were those with an body type that matched the type of leg grouping most commonly seen in training, i.e., '1-3' legs with a rounded shape, or '2-2' legs with an elongated shape. Contradictory stimuli displayed the opposite association, i.e., '1-3' legs with an elongated body shape or '2-2' legs with a rounded body shape. Neutral stimuli possessed upper-bodies that were neither rounded nor elongated. For contradictory stimuli the rule information (i.e., the type of leg grouping) perfectly predicted the stimulus category, but the statistical similarity information (i.e., the shape of the body) predicted the other category. The authors found that participants' reaction times were significantly longer to contradictory stimuli than to congruent stimuli. This result implied that, despite the existence of a perfect rule, participants still take into account statistical information when classifying a stimulus. That is, in categorization, rule knowledge and statistical knowledge interact.

We simulated this experiment as a test of the network's ability to model the interaction of rule and statistical information with reaction times. In addition, an examination of the model's mechanism generates a novel prediction.

A. Training Stimuli

We constructed stimuli according the category structure of Thibaut et al. (1998), which is schematized in Table 3. In the first training phase, all stimuli possessed a leg grouping that defined their category membership ('1-3' or '2-2'); 60% of stimuli in each category possessed one type of body shape (elongated or rounded), 30% of stimuli in each category possessed a neutral body shape, and 10% of stimuli in each category possessed the other body shape.

	'1-3'	*'2-2'*	*Elongated*	*Rounded*	*%*
Cat A	1	0	0	1	60
	1	0	½	½	30
	1	0	1	0	10
Cat B	0	1	1	0	60
	0	1	½	½	30
	0	1	0	1	10

Table 3. Templates for the generation of stimuli in the first phase of Simulation 2. Where 1 indicates the feature was 'on', 0 indicates it was 'off' and ½ indicates it was half on.

In the second, 'associative' training phase, all stimuli possessed a category-defining leg grouping along with the body shape that had appeared with that leg grouping 60% of the time during the first phase, as shown in Table 4.

As in simulation 1, we used stimuli with 10 input elements. We defined one element as corresponding to the feature '1-3 leg grouping', another element as '2-2 leg grouping', a third element as 'rounded body shape' and fourth feature as 'elongated body shape'. For each 'on' feature in a stimulus, we set the pre-normalization value of that feature element to a number randomly chosen from the range 0.9 to 1; when a feature was 'off' it took a pre-normalization value between 0 and 0.2; in instances of neutral body shape, we set both body shape features to an intermediate pre-normalization value, between 0.5 and 0.6.

	'1-3'	*'2-2'*	*Elongated*	*Rounded*
Cat A	1	0	0	1
Cat B	0	1	1	0

Table 4. Templates for the generation of stimuli in the association phase of Simulation 2. All stimuli in a given category took the same form, which was the same template as 60% of stimuli from that category in the first phase.

B. Training Procedure

We trained networks with the same general seven step procedure as in Simulation 1. In the first phase of training, we used 800 exemplars of training stimuli from each category, with the proportions of each type of stimulus within each category reflecting those used by Thibaut et al. (1998). Stimuli were presented in random order.

The first phase was followed immediately by the second 'association phase' of training from Thibaut et al. (1998), in which we presented, in random order, 60 exemplars from each category, constructed according to Table 4.

C. Test Procedure and Stimuli

Following training, we tested networks with congruent, incongruent and neutral stimuli, as in Thibaut et al. (1998). Test stimuli were newly generated from the same templates shown in Table 3, with the 60% stimulus templates from each category corresponding to 'congruent' stimuli, the 30% templates corresponding to 'neutral' stimuli and the 10% templates corresponding to 'incongruent' stimuli. We recorded both categorization accuracy and reaction time.

E. Results and Discussion

Categorization accuracy was not reported by Thibaut et al. (1998), but was high (Thibaut, personal communication), and therefore reaction time was the dependent variable of interest. In our simulations, accuracy on congruent and neutral stimuli was at 100% and accuracy on incongruent stimuli dropped marginally to 98.33%. Our reaction time data showed the same pattern as those of Thibaut et al.: incongruent stimuli took longer to classify than congruent stimuli, as shown in Fig. 9. (Reaction times to neutral stimuli were not discussed by Thibaut et al.; in our experiment, they were comparable to those for congruent stimuli). We performed a repeated measured one-way ANOVA to compare the mean reaction times for congruent, incongruent and neutral stimuli. There was a significant effect of test stimulus type ($F(2,10) = 49.0$, $p<0.001$). Pairwise group comparisons with Sidak adjustment for multiple comparisons revealed that incongruent stimuli took longer to classify than either congruent ($p<0.005$) or neutral ($p<0.005$), but that reaction times for the latter two did not differ from each other ($p=0.95$).

Participants in the study of Thibaut et al. (1998) were interviewed following testing about their awareness of the stimulus attributes possessed by each category. Thirteen out of thirty-three participants noticed the association between the perfectly predictive rule for category membership ('1-3' vs. '2-2'), and the highly correlated dimension ('rounded' vs. 'elongated', respectively). It seems likely that the association between the rule and the correlated dimension was mediated via an association of both with the category identity. The other twenty participants stated that they did not notice any association between '1-3' grouped legs and a rounded body or between '2-2' grouped legs and an elongated body. This finding is extremely interesting because it maps onto a property of the neural networks that emerged during training. Three out of the six networks we trained extracted a weak rule for at least one of the highly correlated (but imperfectly predictive) dimensions, namely 'rounded' or 'elongated'. In each case, the weak rule linked that correlated dimension with the same category as the perfectly predictive, strongly-extracted rule. In other words, if we equate presence of a feature in the rule weights with participants' awareness, these three networks were 'aware' of an association between the highly correlated dimension and the rule, since both were linked to the same category at output. The other three networks did not extract any rules for dimensions that were merely correlated rather than perfectly predictive, mirroring the participants who did not notice the association. In addition, we noticed that the associative phase of training was critical to the emergence of weak rules for the correlated dimensions: in a simulation that we ran *without* the associative phase, only one network out of six extracted any kind of rule for a stimulus dimension that was merely correlated (rather than perfectly predictive) and that rule was so weak as to have negligible influence. Thus, rules for correlated dimensions were extracted only once those correlated dimensions became perfect predictors, in the associative phase. This result makes a prediction for future work: if the Thibaut et al. (1998) study were rerun without the associative phase, no or very few participants should notice an association between the perfect rule and the correlated dimension.

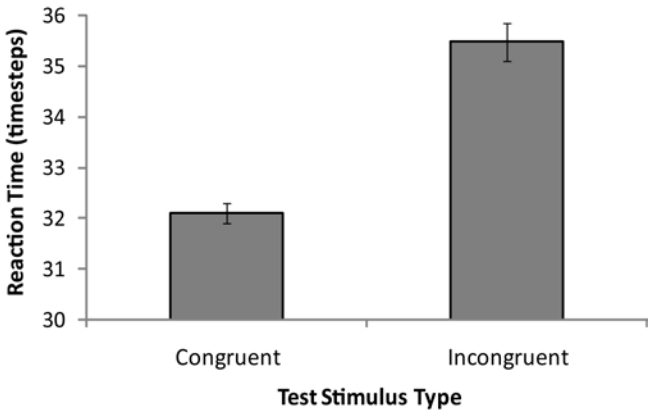

Figure 9. Reaction time ±SEM to congruent and incongruent stimuli in simulations of Thibaut et al. (1998). Only correct responses are included. The unit of reaction time is the number of timesteps in the evolution of output activations until the activation of any node passes threshold.

XIII. GENERAL DISCUSSION

Our model is designed to extract rules that can be described as perfectly predictive of category membership: it learns rule weights for those features that are possessed by members of only a single category and it explicitly suppresses rule weights for those features that appear in more than one category. It does this by exploiting noise in the system to discover which features are category-diagnostic. We demonstrated this mechanism in Simulation 1, in which an ambiguous test stimulus possessing a rule feature from one category and a non-diagnostic feature that had appeared in other categories was correctly classified by networks that possessed the rule mechanism and incorrectly classified by networks that did not. In addition, Simulation 2 further demonstrated this mechanism through the finding that networks did not extract a rule for a feature that was merely highly correlated with category membership (when the simulation was run *without* the 'associative phase' of Thibaut et al., 1998) but that, if a period of training was added during which the correlated feature became perfectly predictive, networks were more likely to extract a rule (three out of six networks did so when the 'associative phase' was simulated). In general, as the period of training including perfectly predictive features lengthens, the

probability of networks extracting a rule increases.

The network employs Hebbian learning, and is a variant upon a physiologically plausible implementation of Kohonen's self-organizing map (Kohonen, 1993). However, the network differs from a standard Kohonen network in two key ways: first, feedback may be provided to the output units such that learning may be supervised or semi-supervised (when feedback is provided on every trial or intermittently, respectively) and, second, an additional set of rule synapses is present, at which learning requires the sequential activation of inputs and outputs by noise.

The network simulates both categorization accuracy and reaction times. In Simulation 2, the model accurately simulated the slowing of reaction times when conflicting information as to category membership was present in a test stimulus. In Simulation 1, the model generated novel predictions for reaction time, along with the predictions for classification accuracy.

XIV. CONCLUSION

We present a connectionist model of category learning, in which there is a mechanism for extracting simple rules for the category membership of stimuli. The mechanism operates by suppressing features that appear in more than one category and selectively focusing on those that are diagnostic for category membership. We claim that rules that emerge in this manner have a distinct status and function with respect to a purely statistical network and, once the rules have emerged, they provide additional power to the categorization system as a whole. We suggest that the mechanisms proposed in this paper provide a plausible manner to bootstrap the development of more complex rules, providing a potential route from an associative, similarity-based system to higher-order, rule-based cognition. Additionally, competition between the associative and rule-based outputs on the category-response layer generates plausible reaction time data. A key novel contribution of this model is the hypothesis that the gradual emergence of rules from simple associative processes is due simply to the presence of noise in the system.

ACKNOWLEDGEMENTS

This work was supported by European Commission Sixth Framework grants FP6-NEST-516542 and FP6-NEST-029088, and by NSF grant BCS-0843773. Thanks to Denis Mareschal and David Huber for insightful suggestions regarding the model, and to Howard Bowman and Emmanuel Pothos for helpful comments on an earlier version of the manuscript.

Appendix: Model Details

Architecture

Feedforward 'statistical' weights are initialized to values drawn from the random distribution between 0 and 0.5. Feedforward 'rule' weights are initialized to zero. All output units possess lateral connectivity (locally excitatory and distally inhibitory). The lateral weights are 'wrap-around', so that the last unit in the row is treated as a neighbor of the first unit in the row. The values of the lateral weights each unit projects are as follows:

Distance to receiving node	*0*	*1*	*2*	*3+*
Value of weight	0.5	0.15	-0.8	-0.5

The size of the output layer is either 9 (Simulation 1) or 8 (Simulation 2), determined by whether there are two to-be-discriminated categories or three. This maintains a relatively constant size of output layer and of category representation, in order that the same lateral weight parameters may serve for both experiments.

Activations

Activation due to stimuli. When a stimulus is presented (whether during training or on test), it is clamped to the input for some time period and output activations are calculated by solving a pair of simultaneous equations – (1) and (2), below – which define the activation of the output units and the activation of a set of inhibitory units that feed into the output units. The equations are solved with the MATLAB function ODE45, which employs the Runge-Kutta method for solving differential equations numerically. The output unit activations are given by:

$$\eta' = w^s \cdot in + w^r \cdot in + w^l \cdot \eta - \xi - c \log\left(\frac{1+\eta}{1-\eta}\right) \quad (1)$$

in which η is a vector of output unit activations, w^s is a matrix containing the statistical weights from input to output units, in is the input pattern, w^l is a matrix containing the lateral weights between output units, c is a constant, ξ is a vector describing the inhibitory units' activations, and the terms in the right hand side of the equation, from left to right, represent: input activation, lateral input from other output units, inhibition from inhibitory units, and leakage. The division in the last term is performed element-wise. The inhibitory unit activations are given by:

$$\xi' = \eta - \theta + b_1\xi - b_2 \log\left(\frac{|1+\xi|}{|1-\xi|}\right) \quad (2)$$

where ξ is a vector of inhibitory unit activations, b_1 and b_2 are constants, θ is an activation threshold, and the terms in the right hand side of the equation, from left to right, represent: activation by the output units, thresholding, recurrent activation from each inhibitory unit to itself, and leakage. The division in the last term is performed element-wise.

When training and testing the full model with presentation of a stimulus (i.e., not noise) output activations are generated via equations (1) and (2). In addition, we run a version of the model in which output activations due to a stimulus are calculated using a modified version of equation (1), in which the rule weights do not influence output activations; this latter case provides a measure of the performance of the statistical weights alone.

Activation due to noise. When noise is presented to the network, it is presented transiently rather than being clamped to the inputs, such that there is assumed to be insufficient duration of input for the full evolution of output unit activation via the differential equations above. Instead, activation due to noise is simply passed to the output units and cycled briefly (7 iterations) through the lateral weights, with the application of a sigmoid function after each iteration. Thus, when noise is presented during training, equations (1) and (2) are replaced by equations (3)–(6).

$$\eta = w^s \cdot in + w^r \cdot in \quad (3)$$

$$\eta_j = \frac{1}{\left(1+e^{-k\left(\eta_j - d\right)}\right)} \text{ for all output nodes } j \quad (4)$$

Then, repeat equations (5) and (6) for 7 iterations:

$$\eta = \frac{1}{2}\eta + \frac{1}{2}w^l \cdot \eta \quad (5)$$

$$\eta_j = \frac{1}{\left(1+e^{-k\left(\eta_j - d\right)}\right)} \text{ for all output nodes } j \quad (6)$$

end

Learning

Statistical Synapses. Learning on the statistical synapses is Hebbian, and is assumed to depend on the simultaneous activation of input and output units. Therefore, statistical synapses are updated when a stimulus is presented, because stimuli are clamped onto the inputs and remain present during evolution of output activations. Statistical synapses are not updated when noise occurs, because noise activation appears only transiently on the inputs and dissipates before the resultant output unit activation has settled. The statistical weight updates are given by:

$$\Delta w_j^s = \eta_j \left(in - w_j^s\left(w_j^s \cdot in\right)\right) \quad (7)$$

$$w_j^s = w_j^s + \alpha \Delta w_j^s \quad (8)$$

where w_j^s is the statistical weight vector from all inputs to output node j, η_j is the activation of output node j, in is the input pattern vector, and α is the Hebbian learning rate. Output node activations are determined by the solution of equations (1) and (2) by the MATLAB function ODE45, which outputs a vector of activations for each output node, corresponding to the activation value at each timestep in the evolution. In equation 3, η_j is the activation at the 'winning timestep', defined as the timestep at which the maximum activation value in the most highly active node was reached.

Feedback in the form of a teaching signal is provided, either intermittently (Simulation 1) or on every trial (Simulation 2). When feedback is provided, the output unit activations in Equation (7) are replaced by a teaching signal, which contains a Gaussian profile of activation across the units designated as the output nodes for that category of stimulus, and zeros elsewhere. The Gaussian profile spans either 3 nodes (for 3 categories, Simulation 1), or 4 nodes (for 2 categories, Simulation 2).

Rule Synapses. Learning on the rule synapses is Hebbian, and is assumed to require non-simultaneous (i.e., sequential) activation of input units and output units. Rule weights are therefore updated only upon activation by noise, and not during activation by a stimulus. Noise activation appears transiently on the input layer, sending activation to the output nodes before dissipating in the input nodes; output activations briefly cycle through the lateral weights before settling. Activations used to update the rule weights (in equations (9) and (10)) are taken from the last iteration of equations (5) and (6). In addition, the rule weights are subject to small decrements, or 'decay', on each trial, in proportion to their size (third term in equation (10)).

$$\Delta w_j^r = \eta_j \left(in - w_j^r \left(w_j^r \cdot in \right) \right) \quad (9)$$

$$w_j^r = w_j^r + \alpha \Delta w_j^r - \tau \left(w_j^r \right)^2 \quad (10)$$

where w_j^r is the rule weight vector from all inputs to output node *j*, η_j is the activation of output node *j*, *in* is the input pattern vector (generated by noise), and τ is the rule decay parameter. In equation (10), the operation $(w_j^r)^2$ is performed element-wise.

No feedback is provided after noise presentation. Learning on the rule weights does not occur during the first 400 training stimuli (the first ¼ to ⅓ of trials), in line with evidence that, early in development, human infants are unable to extract rules (French et al., 2004).

Stimuli and Noise

Input vectors have 10 elements. In training and test stimuli, 'on' features are assigned a value chosen from a uniform random distribution between 0.9 and 1; 'off' features are assigned a value chosen from a uniform random distribution between 0 and 0.2; features designated as 'half on' (Simulation 2) are assigned a value chosen from a uniform random distribution between 0.5 and 0.6. After setting the values of all features, the input vector is normalized.

Noise stimuli are created by setting all features to 0, except a single feature, which is set to 1. That feature is chosen at random from all features that appear among the training stimuli as 'on' or 'half on'.

Training and Test

Stimuli are presented in random order. Each stimulus presentation is followed by update of the statistical synapses. After each training stimulus, three noise bursts appear on the input units, causing update of rule synapses.

Test stimuli are presented at the end of training, with no feedback or learning. The response of the network is taken as being the node that reached the highest activation value at any point in the evolved activations. The response to a test stimulus is judged 'correct' if the most highly active node matches the node that was most often chosen, or second most often chosen, for the category of the test stimulus during the last 120 trials of training. The frequency scores for all nodes for each category accumulate during training by awarding a node 2 points when it is the winning (most highly active) node, and 1 point when it is the runner up.

Reaction Times

RTs are taken as the point at which activation in any node first crosses an activation threshold (in practice, the node crossing first is also the winning node as calculated above). RTs are taken only from trials which yielded correct responses and on which the activation threshold was exceeded.

Parameters

Output Unit Activations: c = 0.4, $b1$ = 0.5, $b2$ = 0.2, θ = 0.3, k = 25, d = 0.3, activation threshold for determining reaction times = 0.7.
Training: α = 0.02, τ = 0.008, proportion of trials supervised = 0.5 (Simulation 1), or 1 (Simulation 2).

REFERENCES

Abraham, C. & Robins A. (2005). Memory retention - the synaptic stability versus plasticity dilemma. *Trends in Neuroscience, 28*(2), 73 – 78.

Allen, S. W. and Brooks, L.R. (1991) Specializing the operation of an explicit rule. *Journal of Experimental Psychology: General*, 120 (1), 3-19.

Ans, B. & Rousset, S. (2000). Neural networks with a self-refreshing memory: knowledge transfer in sequential learning tasks without catastrophic forgetting. *Connection Science, 12(1)*, 1-19 (2000)

Ashby, F. B., Alfonso-Reese, L. A, Turken, A. U., and Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review, 105(3), 442-481.*

Ashby, F. G., Ell, S. W. (2001). The neurobiology of human category learning. *Trends in Cognitive Sciences, 5(5): 204-210.*

Averbeck, B., Latham, P.E., Pouget, A. (2006). Neural correlations, population coding and computation. *Nature Reviews Neuroscience, 7*, 358-366.

Coltheart, M., Curtis, B., Atkins, P., and Haller, M. (1993). Models of reading aloud: Dual route and parallel-distributed-processing approaches. *Psychological Review, 100(4), 589–608.*

Cowell, R. A. and French, R. M. (2007). An unsupervised, dual-network connectionist model of rule emergence in category learning Proceedings of EuroCogSci '07, Taylor & Francis, 318-323.

French, R. M. (1997). Pseudo-recurrent connectionist networks: An approach to the "sensitivity–stability" dilemma. *Connection Science, 9*(4), 353-379.

French, R. M., Mareschal, D., Mermillod, M., & Quinn, P. C. (2004) . The Role of Bottom-up Processing in Perceptual Categorization by 3- to 4-month-old Infants: Simulations and Data. Journal of Experimental Psychology: General, 133, 382-397.

Hebb, D. O. (1949). The Organization of Behavior. NY: John Wiley & Sons.

Keil, F. C. (1989). *Concepts, kinds, and cognitive development.* Cambridge, MA: MIT Press.

Kohonen, T. (1993). Physiological interpretation of the self-organizing map algorithm. *Neural Networks, 6*, 895-905.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review, 99*, 22-44.

Ma, W.J., Beck, J., Latham, P. & Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nature Neuroscience, 9(11)*, 1432-1438.

Marcus, G.F., Vijayan, S., Rao, S.B., Vishton, P.M. (1999). Rule learning in seven-month-old infants. *Science, 283*, 77-80.

O'Reilly, R. & Munakata, Y. (2000) Computational Explorations in Cognitive Neuroscience. Cambridge, MA: The MIT Press.

Pinker, S. & Prince, A. (1988) On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition, 28*, 73-193.

Robins, A. & McCallum, S. (1999). The consolidation of learning during sleep: Comparing the pseudorehearsal and unlearning accounts. *Neural Networks, 12*: 1191 - 1206

Rossi, F. (1996) Attribute suppression with multi-layer perceptron. *Proc.*

IMACS/IEEE Multiconference on Computational Engineering in Systems Applications (CESA'96), Symposium on Robotics and Cybernetics, pp. 542-547, *Lille, France, July 9-12.* [Note(s): 943].

Seidenberg, M. S. and McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review, 96*, 523-568.

Smolensky, P. (1988) On the proper treatment of connectionism. *Behavioral and Brain Sciences, 11,* 1-74.

Spitzer, M. (1999) *The Mind Within the Net: Models of Learning, Thinking and Acting*, Bradford Books.

Thibaut, J.P., Lemaire, F. and Quadri, J. (1998). Categorization under the Influence *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society*, Mahwah, NJ: LEA, 1055-1060.

Thomas E., Van Hulle M. & Vogels R. (2000). Encoding of categories by non-category specific neurons in the inferior temporal cortex. *Journal of Cognitive Neuroscience 13.* 190-200.

Traynelis, S. F. & Jaramillo, F. (1998). Getting the most out of noise in the central nervous system. *Trends in Neuroscience, 21,* 137-145.

Usher, M. and McClelland, J. L. (2001) The time course of perceptual choice: The Leaky, Competing, Accumulator Model. *Psychological Review*, 108 (3), 550-592.