A Knowledge-Resonance (KRES) Model of Category Learning

Bob Rehder (**bob.rehder@nyu.edu**)
Department of Psychology, New York University, 6 Washington Place
New York, NY 10003 USA

Gregory L. Murphy (**gmurphy@s.psych.uiuc.edu**)
Beckman Institute, University of Illinois, 405 N. Mathews Ave
Urbana, IL 61801 USA

Abstract

In this article we present a connectionist model of category learning that takes into account the prior knowledge that people bring to many new learning situations. This model, which we call the *Knowledge-Resonance Model* or *KRES*, employs a recurrent network with bidirectional connections which are updated according to a *contrastive-Hebbian learning rule*. When prior knowledge is incorporated into a KRES network, the KRES activation dynamics and learning procedure accounts for a range of empirical results regarding the effects prior knowledge on category learning, including the accelerated learning that occurs in the presence of knowledge, the reinterpretation of features in light error correcting feedback, and the unlearning of prior knowledge which is inappropriate for a particular category.

A traditional assumption in category learning research is that learning is based on those category members people observe and is relatively independent of the prior knowledge that they already possess. According to this *data-driven* or *empirical learning* view of category learning, people associate observed exemplars and the features they display (or a summary representation of those features such as a prototype or a rule) to the name or label of the category. In this account there is neither need nor room for the influence of the learner's prior knowledge of how those features are related to each other or other concepts on the learning process. In contrast, the last several years has seen a series of empirical studies that demonstrate the dramatic influence that a learner's prior knowledge often has on the learning process in interpreting and relating a category's features to one another and other concepts. Indeed, knowledge effects have been demonstrated in every area of conceptual processing in which they have been investigated (see Murphy, 1993, for a review).

The goal of this article is to introduce a theory of category learning that accounts for the effects of prior knowledge on the learning of new categories. This theory, which we refer to as the *Knowledge-Resonance Model*, or *KRES*, is a connectionist network that specifies prior knowledge in the form of existing concepts and relations between concepts. We will show that when knowledge is incorporated into a KRES network, KRES's activation and learning procedures account for a number of empirical results regarding the effects of prior knowledge on category learning.

Other connectionist models have been proposed to account for the learning of new categories (e.g., Gluck & Bower, 1988; Kruschke, 1992), and these models have generally used feedforward networks (i.e., activation flows only from inputs to outputs) and learning rules based on error signals that traverse the network from outputs to inputs (e.g., backpropagation). KRES departs from these previous models in two regards. First, rather than feedforward networks, KRES uses *recurrent networks* in which activation is allowed to flow not only from inputs to outputs but also from outputs to inputs and back again. Recurrent networks respond to inputs by each unit iteratively adjusting its activation in light of all other units until the network "settles," that is, until change in units' activation levels ceases. This settling process can be understood as an interpretation of the input in light of the knowledge encoded in the network. As applied to the categorization problems considered here, a KRES network accepts inputs that represent an object's features, and interprets (i.e., classifies) that object by settling into a state in which the object's correct category label is active.

Second, rather than backpropagation, KRES employs *contrastive Hebbian learning* (CHL) as a learning rule (Movellan, 1989; O'Reilly, 1996). Backpropagation has been criticized for being neurally implausible because it assumes nonlocal information regarding the error generated from corrective feedback in order for connection weights to be updated. In contrast, CHL transmits error by using the same connections that propagate activation. During an initial *minus phase*, a network is allowed to settle in light of an input pattern. In the ensuing *plus phase*, the network is provided with what serves as error-corrective feedback by being presented with the output pattern that should have been computed during the minus phase and allowed to resettle in light of that (correct) output pattern. Connection weights are then updated as a function of the difference between the activation of units in the two phases.

In the following sections we first describe KRES and then present three simulations of human category learning data. We will show how KRES's successes can be attributed to its recurrent network that allows category features to be interpreted in light of prior knowledge, and the CHL learning algorithm that allows (re)learning of all connections in a network, including those that represent prior knowledge.

The Knowledge-Resonance Model (KRES)

An example of a KRES model is presented in Figure 1. In Figure 1, circles depict *units* that represent concepts that are

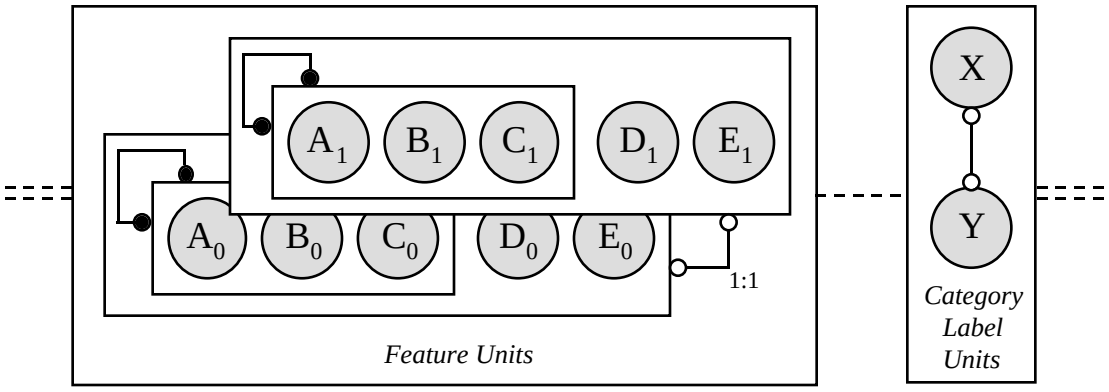

Figure 1. A sample KRES model.

either category labels (X and Y) or features (A_0, A_1, B_0, B_1, C_0, C_1, etc.). To simplify the depiction of connections among units, units are organized into *layers* specified by rectangles. Solid lines between layers represent connections among units. Solid lines terminated with black circles are excitatory connections, whereas those terminated with hollow circles are inhibitory connection. Dashed lines represent new, to-be-learned connections. Two connected layers are fully connected (i.e., every unit is connected to every other unit), unless annotated with "1:1" (i.e. "one-to-one") in which case a unit in a layer is connected to only one unit in the other layer. Finally, double dashed lines represent sources of external inputs. As described below, both the feature units and the category label units receive external input, although at different phases of the learning process.

We now describe the basic elements of KRES, which include its representation assumptions, activation dynamics (i.e., constraint satisfaction), and learning via CHL.

Representational Assumptions

At any time a unit has a level of activation in the range 0 to 1 that represents the activation of the concept. A unit *i*'s activation act_i is a sigmoid function of its total input,

$$act_i = 1 / [1+ \exp (-total\text{-}input_i)]$$

and its total input comes from three sources,

$$total\text{-}input_i = net\text{-}input_i + external\text{-}input_i + bias_i.$$

Network input represents the input received from other units. External input represents the presence of (evidence for) the concept in the external environment. Finally, a unit's bias can be interpreted as a measure of the prior probability that the concept is present in the environment.

In many applications, two or more features might be treated as mutually exclusive values on a single dimension. In Figure 1 the stimulus space is assumed to consist of five binary valued dimensions, with A_0 and A_1 representing the values on dimension A, B_0 and B_1 the values on dimension B, etc. To represent that these feature pairs are mutually exclusive they are linked by inhibitory connections. The category labels X and Y are also assumed to be mutually exclusive and are linked by an inhibitory connection.

Connections between units are symmetric, that is, $weight_{ij} = weight_{ji}$. A unit's network input is computed by multiplying the activation of each unit to which it is connected by the connection's weight, and then summing over those units in the usual manner,

$$net\text{-}input_i = \sum_j act_i * weight_{ij} .$$

KRES primarily represents prior knowledge in the form of prior relations between features. For example, in Figure 1 it is assumed that features A_0, B_0, and C_0 are related by prior knowledge, as are features A_1, B_1, and C_1. These relations are rendered as excitatory connections between the features. In KRES prior knowledge can also be represented in the form of preexisting concepts (i.e., units) and excitatory connections that link those preexisting concepts to the feature units (see Simulation 3 below).

Classification via Constraint Satisfaction

Before a KRES model is presented with input that represents an object's features, the activation of each unit is initialized to a value determined solely by its bias. The external input of a feature unit is then set to 1 if the feature is present in the input, -1 if it is absent, and 0 if its presence or absence is unknown. The external input of all other units is set to 0. The model then undergoes a multi-cycle constraint satisfaction processes which involves updating the activation of each unit in each cycle in light of its external input, its bias, and its current network input. (In each cycle, the serial order of updating units is determined by randomly sampling units without replacement.) After each cycle the *harmony* of the network is computed, given by,

$$harmony = \sum_i \sum_j act_i * act_j * weight_{ij} . \quad (1)$$

Constraint satisfaction continues until the network settles, as indicated by a change in harmony from one cycle to the next of less than 0.00001.

The activation of units X and Y that result from this settling process represent the evidence that the current input should be classified as an X or Y, respectively. These activation values can be mapped into a categorization decision in the standard way, that is, according to Luce's choice axiom,

$$choice\text{-}probability\ (X, Y) = act_X / (act_X + act_Y) .$$

Contrastive Hebbian Learning (CHL)

As described earlier, the settling of a network that results from presenting just the feature units with input is referred to as the minus-phase. In the plus-phase, error-correcting feedback is provided to the network by setting the external

inputs of the correct and incorrect category label units to 1 and –1, respectively, and allowing the network to resettle in light of these additional inputs. We refer to the activation values of unit *i* that obtain after the minus and plus phases as act_i^- and act_i^+, respectively. After the plus phase the connection weights are updated according to the rule,

$$\Delta weight_{ij} = lrate * (act_i^+ * act_j^+ - act_i^- * act_j^-) \qquad (2)$$

where *lrate* is a learning rate parameter.

Network Training

Before training a KRES network, all connections weights are set to their initial values. In the following simulations, all to-be-learned connections are initialized to a random value in the range [-0.1, 0.1], and the biases of all units are initialized to 0. As in the behavioral experiments we simulate, training consists of repeatedly presenting a set of training patterns in blocks with the order of the patterns randomized within block. Training continues until the error for a block falls below an error criterion of 0.10. The error for a block is computed by summing the errors associated with each training pattern in the block and dividing by the number of patterns. The error associated with a training pattern is the sum of the squared differences between the activation levels of the category label units and their correct values (0 or 1).

KRES Simulation of Empirical Data

We present KRES simulations of three empirical data sets that illustrate the effect of prior knowledge on category learning. The KRES model was rerun ten times with a different set of random weights, and the results reported below are averaged over those ten runs.

Simulation 1: Murphy and Allopenna (1994)

In the literature on category learning with prior knowledge, perhaps the most pervasive effect is that learning is dramatically accelerated when the prior knowledge is consistent with the empirical structure of training exemplars. For example, Murphy and Allopenna (1994, Experiment 2), presented examples of two categories the features of which either could (Theme Condition) or could not (No Theme Condition) be related to one another. In the Theme condition one category had six typical features that could be related because they could be construed as features of arctic vehicles ("drives on glaciers," "made in Norway," "heavily insulated," etc.) whereas the other category had six typical features that could be construed as features of jungle vehicles ("drives in jungles," "made in Africa," "lightly insulated," etc.). In the No Theme condition, the typical features of the categories could not be related to one another. Exemplars also possessed three knowledge-irrelevant features which were not predictive of category membership. Murphy and Allopenna found that participants reached a learning criterion in fewer blocks in the Theme (2.5) versus the No Theme condition (4.1), a result the authors attribute to the knowledge relating the features in the Theme condition.

This experiment was simulated by a KRES model like the one shown in Figure 1 with 18 features representing the two values on 9 binary dimensions. In the Theme but not the No Theme condition the six related features in each of the two

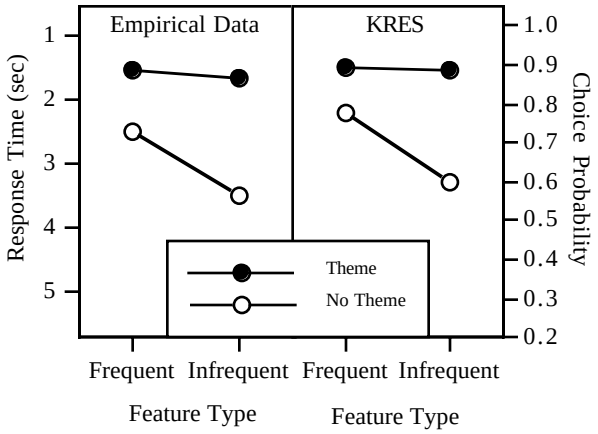

Figure 2. Results from Murphy & Allopenna (1994).

categories were linked with excitatory connections. The weight on these excitatory connections was initialized to 0.4, the inhibitory connections were initialized to –2.0, and the learning rate was set to 0.10.

The results indicated that KRES reproduces the learning advantage found in the Theme condition: The error criterion was reached in fewer blocks as compared to the No Theme condition (2.0 vs. 4.0). This advantage can be attributed to KRES's use of recurrent networks: The mutual excitation of knowledge-relevant features in the Theme condition resulted in higher activation values for those units, which in turn led to the faster growth of the connection weights between the features and category label units (according to the CHL learning rule Eq. 2). Once some learning of those connections has occurred, the higher activation of the features also leads to greater activation of the category labels themselves.

Murphy and Allopenna also varied the frequency with which the six knowledge-relevant features appeared during training, and then tested how subjects classified those features during an ensuing test phase. The left side of Figure 2 indicates that, as expected, RTs for these single-feature classification trials were shorter for frequent versus infrequent features in the No Theme condition. In contrast, in the Theme condition RTs were insensitive to features' empirical frequency. This pattern of results was also reflected in subjects' categorization accuracy. (Note Figure 2's RT scale has been inverted to facilitate comparison with KRES's choice probabilities presented below.)

To determine whether KRES would also exhibit these effects, after training the model was presented with single features. That is, the unit representing that feature was given an external input of 1, the unit representing the other feature on the same dimension was given an input of –1, and all other units were given an input of 0. The right side of Figure 2 indicates that KRES's choice probabilities reproduce the pattern of results for the single-feature tests. In KRES, infrequently presented knowledge-relevant features are classified nearly as accurately as frequently presented ones because during training those features were activated by inter-feature excitatory connections even on trials in which they were not presented, and hence were associated with the category label nearly as strongly as knowledge-relevant features that were frequently presented.

Simulation 2: Kaplan and Murphy (2000)

Simulation 1 provides evidence in favor of KRES's use of recurrent networks to accelerate learning by amplifying the activation of features interconnected by prior knowledge. However, another distinctive characteristic of KRES is that the category label units are also recurrently connected to the features. In this section we provide evidence that activation also flows backwards from category label units.

Using a modified version of the materials used in Murphy and Allopenna (1994), Kaplan and Murphy (2000, Experiment 4) provided an especially dramatic demonstration of the effect of prior knowledge. In that study, participants were presented with training examples that contained only *one* of the knowledge-relevant features and up to six knowledge-irrelevant features that were predictive of category membership. That is, the single knowledge-relevant feature in each exemplar had prior associations only to features in other category exemplars. Under these conditions, one might have predicted that participants would be unlikely to notice the relations among the features in different exemplars, especially given that those features were each embedded in an exemplar with many knowledge-irrelevant features. In fact, participants in this Intact Theme condition reached a learning criterion in fewer blocks (2.7) than did those in a No Theme condition (5.0) in which the categories had the same empirical structure but no relations among features.

We simulated this experiment with a KRES model with 22 features on 11 binary dimensions. In the Intact Theme condition the features within the two sets of six knowledge-relevant features were inter-related with excitatory connections, as in Simulation 1. The weight on these excitatory connections was initialized to 0.35, the inhibitory connections were set to –2.0, and the learning rate was set to 0.10.

KRES reproduced the learning advantage found in the Intact Theme condition (3.0 blocks) as compared to the No Theme condition (5.4). This advantage obtained because even though each training pattern in the Intact Theme condition contained only one knowledge-relevant feature, that feature tended to activate the knowledge-relevant features to which it was connected, and hence the connections between each knowledge-relevant feature and its correct category label were strengthened on every trial to at least some degree.

After each training block, Kaplan and Murphy also presented test blocks in which participants classified each of the 22 features. The left side of Figure 3 indicates that as expected after the final block of training participants in the No Theme condition were faster at classifying those features that appeared in several training exemplars (Characteristic features) than those that appeared in just one (Idiosyncratic features). In contrast, in the Intact Theme condition participants were faster at classifying the Idiosyncratic features, because they were also knowledge-relevant. Unexpectedly, Intact Theme participants were also faster at classifying the Characteristic features (i.e., the knowledge-irrelevant features) even though those features were not related via prior knowledge, and even though Intact Theme participants had experienced fewer training blocks on average (2.7 vs. 5.0).

This latter result is a challenge for many standard connectionist accounts of learning, because in such accounts the better learning associated with knowledge-relevant features would be expected to *overshadow* the learning of knowledge-irrelevant features (Gluck & Bower, 1988)–that is, these features should be worse with knowledge than without as a result of them competing with the stronger knowledge-relevant features. In contrast, Figure 3 indicates that KRES is able to account for the better learning (or in some experiments, equal learning) of the knowledge-irrelevant features in the Intact Theme condition. This result can be attributed to the use of recurrent connections to the category label units. After some excitatory connections between the knowledge-irrelevant features and category labels have been formed, the knowledge-relevant and -irrelevant features began to activate each other through the category node. This greater activation of the knowledge-irrelevant features leads to accelerated learning of their connection weights to the category labels. That is, KRES's use of recurrent networks compensates for the effects of cue competition found in the usual feedforward network.

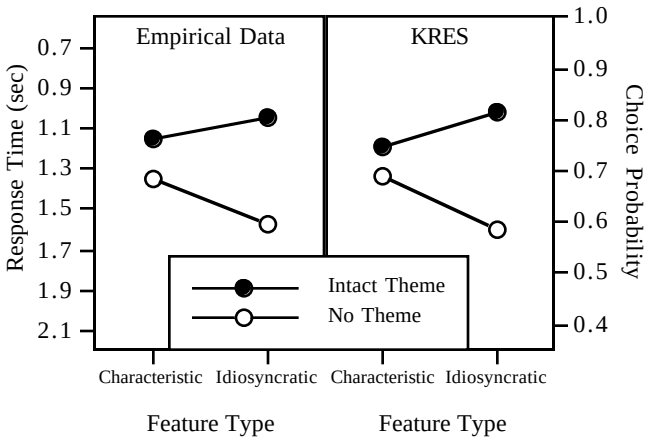

Figure 3. Results from Kaplan & Murphy (2000). In the Intact Theme condition Idiosyncratic features are knowledge-relevant and Characteristic features are knowledge-irrelevant.

Simulation 3: Wisniewski and Medin (1994)

In a final simulation we demonstrate the efficacy of contrastive-Hebbian learning to update weights on connections not involving the category label units. In particular, we examine KRES's ability to update connections representing prior knowledge that is inappropriate in the current context.

Wisniewski and Medin (1994, Experiment 2) present empirical results that call into question the assumption of standard theories of category learning that features can be identified prior to learning. Participants were shown two categories of line drawings of persons that were described as drawn by *creative* and *non-creative children* or by *farm* and *city kids*. Wisniewski and Medin chose to use line drawings to illustrate that what constitutes a feature in a stimulus depends on the prior expectations that one has about its possible category membership. For example, they found that participants would assume the presence of *abstract features* about a category depending on the category's label (e.g., creative children's drawings depict unusual amounts of detail and characters performing actions) and examine the drawings for concrete evidence of those abstract features in order to

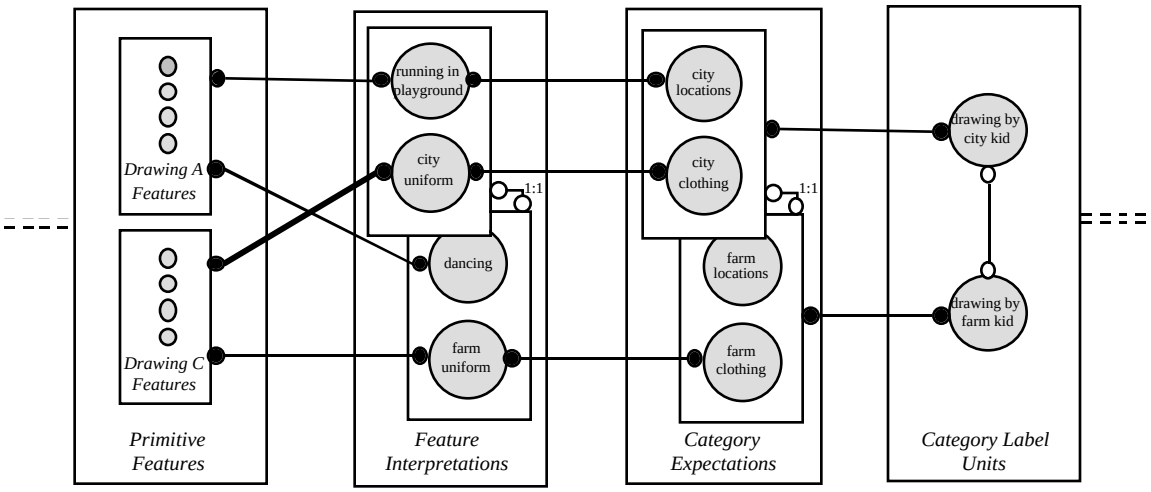

Figure 4. KRES model for Simulation 3.

determine its category membership. They also found that the feedback participants received about category membership led them to change their original interpretation of certain features of the line drawings. For example, after first interpreting a character's clothing as a farm "uniform" (and categorizing the picture as drawn by a farm kid), some participants reinterpreted the clothing as a city uniform after receiving feedback that the picture was drawn by a city kid.

To demonstrate these effects with KRES, we imagined a simplified version of the materials of Wisniewski and Medin's in which there were only two drawings. One drawing (Drawing A), was of a character performing an action interpretable either as climbing in a playground or dancing. In the other (Drawing C), a character's clothing could be seen as a farm uniform or a city uniform. These alternative interpretations are represented in the left side of the KRES model of Figure 4. Whereas we assume the two interpretations of Drawing A are equally likely, we assume that a city uniform is the more likely interpretation of Drawing C (as depicted by the heavier line connecting the features of Drawing C and their city uniform interpretation). The alternative interpretations are connected with inhibitory connections representing that only one interpretation is correct.

The model of Figure 4 was presented with the problem of learning to classify Drawing A as done by a city kid, and Drawing C by a farm kid. We represented the expectations or hypotheses that Wisniewski and Medin found that learners form in the presence of meaningful category labels such as *farm* or *city kids* as units connected via excitatory connections to the category labels, as shown in the right side of Figure 4. In Figure 4, city and farm kids are expected to be in locations and wear clothing appropriate to cities and farms. These expectations are in turn related by excitatory connections to the picture interpretations that instantiate them: climbing in a playground instantiates a city location, and city and farm uniforms instantiate city and farm clothing, respectively. In Figure 4, all inhibitory connections were set to –3.0 and all excitatory connections were set to 0.25, except for those between Drawing C's features and their city uniform interpretation, which were set to 0.30.

Before a single training trial is conducted, the prior knowledge incorporated into this KRES model is able to decide on a classification of both drawings. Upon presentation of Drawing A, its two interpretations, climbing-in-a-playground or dancing are activated, and climbing-in-a-playground in turn activates the city location expectation, which in turn activates the category label for city kids' drawings. The drawing is correctly classified as having been drawn by a city kid. Moreover, as the network continues to settle, activation is sent back from the category label to the climbing-in-a-playground unit. As a result, the climbing-in-a-playground interpretation of Drawing A is more active than the dancing interpretation when the network settles. That is, the top-down knowledge provided to the network results in the resolution of an ambiguous feature (i.e., the action is interpreted as climbing in a playground rather than dancing). Wisniewski and Medin found that the same drawing would be interpreted as depicting dancing instead when participants were required to classify the drawings as having been done by creative or noncreative children.

Similarly, upon presentation of Drawing C, its two interpretations are activated, but because the city uniform interpretation receives more input as a result of its larger connection weight, it quickly dominates the farm uniform interpretation. As a result, the category label for city kids' drawings becomes active (via the city clothing expectation). That is, the drawing is *incorrectly* classified as having been drawn by a city kid. However, error feedback results in the model changing its interpretation of Drawing C. During the model's plus phase, the farm kids' category label is more active than the city kids' label as a result of the external inputs those units receive. The activation emanating from the farm kids' label leads to the activation of the farm clothing expectation and then the farm uniform feature interpretation, which ends up dominating the city uniform unit.

This result indicates that KRES can reinterpret features in light of error feedback. The more important question, however, is whether KRES can *learn* this new interpretation so that Picture C (or a similar picture) will be correctly classified in the future. The left side of Figure 5 shows the changes to the connection weights brought about by the CHL learning rule with a learning rate of 0.3 as a function

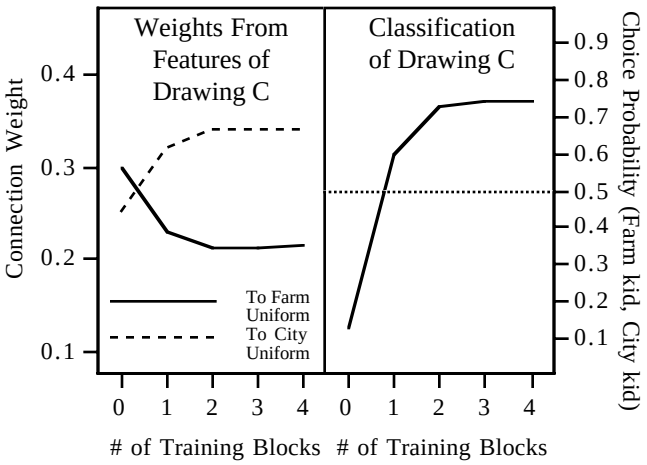

Figure 5. Results from Simulation 3.

of number of blocks of training on the two drawings. Figure 5 indicates that the connection weights associated with the interpretation of Drawing C as a city uniform rapidly decrease from their starting value of 0.30, while the weights associated with Drawing C's interpretation as a farm uniform increase from their starting value of 0.25. As a result, after just one training block KRES's classification of Drawing C switches from being done by a city kid to a farm kid (as indicated by the choice probabilities shown in the right side of Figure 5). That is, KRES uses the error feedback it receives to learn a new interpretation of Drawing C.

General Discussion

We have presented a new model of category learning that attempts to account for the influence of prior knowledge that learners often bring to the task of learning a new category. KRES utilizes a recurrent network in which knowledge is encoded in the form of connections among units. We have shown the changes brought about by this recurrently-connected knowledge to the interpretations and reinterpretations of a category's features provides a reasonable account of three data sets exhibiting the effects of prior knowledge on category learning. In Simulation 1 we demonstrated how KRES's recurrent network provides a pattern of activation among features that accounts for the finding that knowledge accelerates the learning of connections to category labels. In Simulation 2 we demonstrated that the presence of knowledge does not inhibit the learning of knowledge-irrelevant features, a striking result in light of well-known learning phenomena such as cue competition. In Simulation 3 top-down flow of activation was instrumental in KRES's success in resolving the ambiguity surrounding the interpretation of a perceptual features. Moreover, the CHL learning rule allowed the knowledge responsible for one interpretation of an ambiguous feature to be unlearned and a new interpretation learned when the network was provided with feedback regarding the stimulus's correct category.

KRES departs from previous connectionist models that attempt to account for the effects prior knowledge with feedforward networks. For example, Heit & Bott (2000) have proposed a model, *Baywatch*, that assumes that features send activation to prior concepts, that both the features and the prior concepts send activation to the category label units, and that learning consists of learning the connections to the category labels. Although we believe that existing categories often aid the learning of new categories (e.g., our knowledge of VCRs helps us understand DVD players), the Baywatch approach is limited to the learning of new categories that are essentially refinements of existing concepts. In contrast, KRES only assumes the presence of relations between features to account for the data in Simulations 1 and 2, and hence is able to learn truly new concepts, not just refinements of existing ones.

There remains much to be discovered about the properties of recurrent networks and contrastive Hebbian learning with regard to the learning of categories. However, we believe that recurrent networks are likely to be critical to any attempt at accounting for the effects of prior knowledge on category learning. For example, standard feedforward networks seem intrinsically unable to account for (a) the accelerated learning produced by prior knowledge without presupposing prior knowledge of the to-be-learned category, (b) the effects of top-down knowledge on resolving ambiguous features, and (c) the reinterpretation of ambiguous features in light of feedback regarding category membership.

Acknowledgements

This work was supported by NSF Grant SBR 97-20304.

References

Gluck, M. A., & Bower, G. H. (1988). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General, 117*, 227-247.

Heit, E., & Bott, L. (2000). Knowledge selection in category learning. In D. L. Medin (Ed.), *The psychology of learning and motivation.* (pp. 163-199). Academic Press.

Kaplan, A. S., & Murphy, G. L. (2000). Category learning with minimal prior knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*, 829-846.

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

Movellan, J. R. (1989). Contrastive Hebbian learning in the continuous Hopfield model. In D. S. Touretzky, G. E. Hinton, & T. J. Sejnowski (Eds.), *Proceedings of the 1989 Connectionist Models Summer School* .

Murphy, G. L. (1993). Theories and concept formation. In I. V. Mechelen, J. Hampton, R. Michalski, & P. Theuns (Eds.), *Categories and concepts: Theoretical views and inductive data analysis.* (pp. 173-200). Academic Press.

Murphy, G. L., & Allopenna, P. D. (1994). The locus of knowledge effects in concept learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*, 904-919.

O'Reilly, R. C. (1996). Biologically plausible error-driven learning using local activation differences: The generalized recirculation algorithm. *Neural Computation, 8*, 895-938.

Wisniewski, E. J., & Medin, D. L. (1994). On the interaction of theory and data in concept learning. *Cognitive Science, 18*, 221-282.