Learning Involves Attention

John K. Kruschke
Dept. of Psychology
1101 E. 10th St.
Indiana University
Bloomington, IN 47405-7007
kruschke@indiana.edu
(812) 855-3192

Chapter to appear in:
G. Houghton (Ed.), *Connectionist Models in Cognitive Psychology*. London: Psychology Press.
Preliminary draft version of February 6, 2000. Comments welcome. Please do not quote.

Abstract

Attention in learning

One of the primary factors in the resurgence of connectionist modeling is these models' ability to learn input-output mappings. Simply by presenting the models with examples of inputs and the corresponding outputs, the models can learn to reproduce the examples and to generalize in interesting ways. After the limitations of perceptron learning (Minsky & Papert, 1969; Rosenblatt, 1958) were overcome, most notably by the backpropagation algorithm (Rumelhart, Hinton, & Williams, 1986) but also by other ingenious learning methods (e.g., Ackley, Hinton, & Sejnowski, 1985; Hopfield, 1982), connectionist learning models exploded into popularity. Connectionist models provide a rich language in which to express theories of associative learning. Architectures and learning rules abound, all waiting to be explored and tested for their ability to account for learning by humans or other animals.

A thesis of this chapter is that connectionist learning models must incorporate rapidly shifting selective attention and the ability to learn attentional redistributions. This kind of attentional shifting is not only necessary to mimic learning by humans and other animals, it is also a highly effective and rational solution to the demands of learning many new associations as quickly as possible. This chapter describes three experiments (one previously published and two new) that demonstrate the action of attentional learning. All the results are fit by connectionist models that shift and learn attention, but the results cannot be fit when the attention mechanisms are shut off.

Shifts of attention facilitate learning

A basic fact of learning is that people quickly learn new associations without rapidly forgetting old associations. Presumably this ability is highly adaptive for any creature that confronts a rich and complex environment. Consider a hypothetical situation in which an animal learns that mushrooms with a round top and smooth texture are tasty and nutritious. After successfully using this knowledge for some time, the animal encounters a new mushroom with a smooth texture but a flat top. This mushroom turns out to induce nausea. How is the animal to quickly learn about this new kind of mushroom, without destroying still-useful knowledge about the old kind of mushroom? If the animal learns to associate both features of the new mushroom with nausea, then it will inappropriately destroy part of its previous knowledge about healthy mushrooms. That is, the previous association from smooth texture to edibility will be destroyed. On the other hand, if the old association is retained, it generates a conflicting response (i.e., eating the mushroom).

To facilitate learning about the new case, it would be advantageous to selectively attend to the distinctive feature, namely, flat top, and learn to associate this feature with nausea. By selectively attending to the distinctive feature, previous knowledge is preserved, and new learning is facilitated. Not only should attention be shifted in this way to facilitate learning, but the shifted attentional distribution should itself be learned: Whenever the animal encounters a mushroom with smooth texture and flat top, it should shift attention to the flat top, away from the smooth texture. This will allow the animal to properly anticipate nausea, and to avoid the mushroom.[1] The third example in this chap-

[1] An alternative possible solution would be to encode the entire configuration of features in each type of mushroom, and to disallow any generalization on the basis of partially matched configurations. In this way, knowledge about smooth

ter describes a situation in which people use exactly this kind of attentional shifting during learning. The challenge to the theorist is expressing these intuitions about attention in a fully specified model.

Shifts of attention can be assessed by subsequent learning

The term “attention,” as used here, refers to both the influence of a feature on an immediate response and the influence of a feature on learning. If a feature is being strongly attended to, then that feature should have a strong influence on the immediate response and on the imminent learning. This latter influence of attention on learning is sometimes referred to as the feature’s *associability*. In this chapter, these two influences of attention are treated synonymously. This treatment is a natural consequence of the connectionist models described below, but the treatment might ultimately turn out to be inappropriate in the face of future data.

Because redistribution of attention is a learned response to stimuli, the degree of attentional learning can be assayed by examining *subsequent* learning ability. If a person has learned that a particular feature is highly indicative of an appropriate response, then, presumably, the person has also learned to attend to that feature. If subsequent training makes a different feature relevant to new responses, then learning about this new correspondence should be relatively slow, because the person will have to unlearn the attention given to the now-irrelevant feature. In general, learned attention to features or dimensions can be inferred from the ease with which subsequent associations are learned. This technique is used in all three examples presented below.

Intra- and extra-dimensional shifts

A traditional learning paradigm in psychology investigates perseveration of learned attention across phases of training. In the first phase, participants learn that one stimulus dimension is relevant to the outcome while other dimensions are irrelevant. In the second phase, the mapping of stimuli to outcomes changes so that either a different dimension is relevant or the same dimension remains relevant. The former change of relevance is called *extra*dimensional shift, and the latter change is called *intra*dimensional shift. Many studies in many species have shown that intradimensional shift is easier than extradimensional shift, a fact that can be explained by the hypothesis that subjects learn to attend to the relevant dimension, and this attentional shift perseverates into the second phase (e.g., Mackintosh, 1965; Wolff, 1967). In this section of the chapter, a recent experiment demonstrating this difference is summarized, and a connectionist model that incorporates attentional learning is shown to fit the data, whereas the model cannot fit the data if its attentional learning mechanism is “turned off.”

Experiment design and results

Consider the simple line drawings of freight train box cars shown in Figure 1. They vary on three binary dimensions: height, door position, and wheel color. In an experiment conducted in my

round mushrooms would not interfere with knowledge about smooth flat mushrooms, despite the fact that both pieces of knowledge include the feature smoothness. A problem with this approach is that knowledge does not generalize from learned cases to novel cases, yet generalization is perhaps the most fundamental goal of learning in the first place. For a discussion of configural and elemental learning, see the chapter by Shanks in this collection.

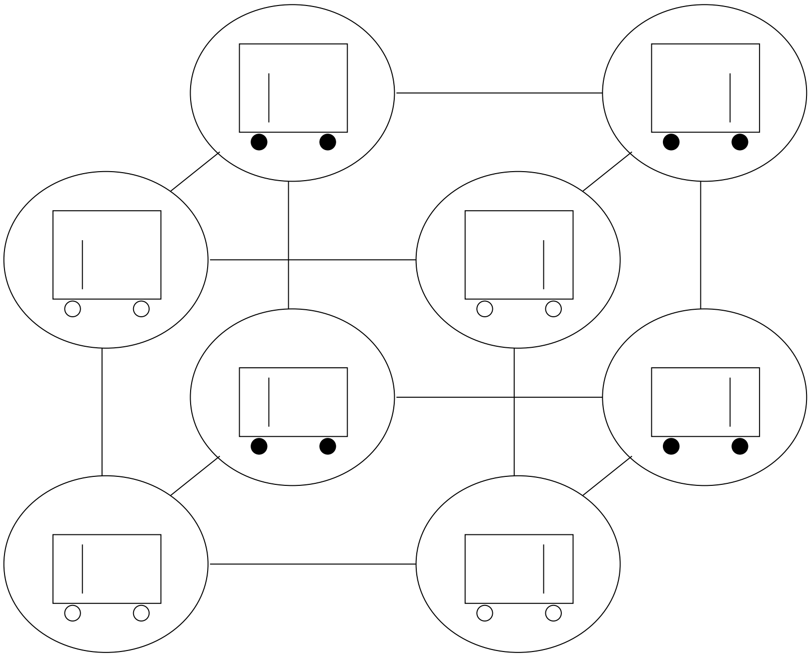

Figure 1. Stimuli used for relevance shift experiment of Kruschke (1996b). (The ovals merely demarcate the different stimuli and are not part of the stimuli per se. The lines connecting the ovals indicate the dimensions of variation between stimuli.)

lab (Kruschke, 1996b), people learned to classify these cars into one of two routes. On each trial in a series, a car would appear on a computer screen, the learner would make his or her choice of the route of the car by pressing a corresponding key, and then the correct route would be displayed. During the first few trials, the learner could only guess, but after many trials, she or he could learn the correct answers.

Figure 2 indicates the mapping of cars to routes. The the cubes in Figure 2 correspond with the cube shown in Figure 1. Each corner is marked with a disk whose color indicates the route taken by the corresponding train; in other words, the color of the disk indicates the category of the stimulus.

The left side of Figure 2 shows the categorization learned in the first phase of training, and the right side shows the categorization learned subsequently. In the first phase, it can be seen that the vertical dimension is irrelevant. This means that variation on the vertical dimension produces no variation in categorization: The vertical dimension can be ignored with no loss in categorization accuracy. The other two dimensions, however, are relevant in the first phase. Some readers might recognize this as the exclusive-or (XOR) structure on the two relevant dimensions.

In the subsequent phase, some learners experienced a change to the top-right structure of Figure 2, and other learners experienced a change to the bottom-right structure. In both of these second-phase structures only one dimension is relevant, but in the top shift this relevant dimension was one of the initially relevant dimensions, so the shift of relevance is called intradimensional, whereas in

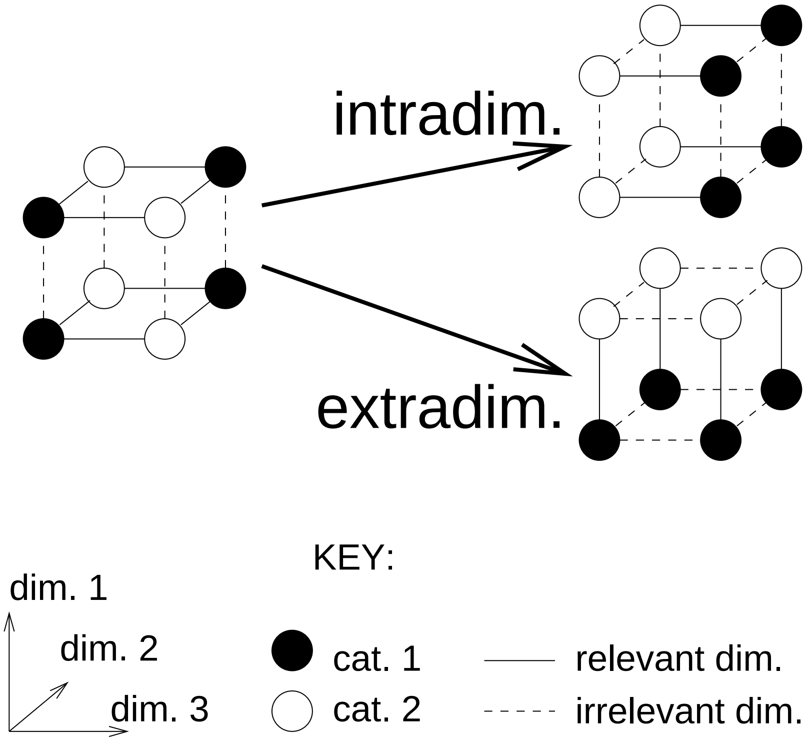

Figure 2. The structure of two types of relevance shifts (adapted from Kruschke, 1996b). The cube at left indicates the initially learned categorization; the cubes at right indicate the alternative subsequently learned categorizations.

the bottom shift the newly relevant dimension was initially irrelevant, so the shift of relevance is called extradimensional. Notice that the two second-phase category structures are isomorphic, so any differences in ease of learning the second phase cannot be attributed to differences in structural complexity.

This design is an advance over all previous studies of shift learning because no novel stimulus values are used in the either shift. Thus intradimensional and extradimensional shifts can be directly compared without confounded changes in novelty. In traditional studies of intradimensional shift, the shift is accompanied by introduction of novel values on the relevant dimension. For example, the initial phase might have color relevant, with green indicating category X and red indicating category Y. The only way to implement an intradimensional shift, without merely reversing the assignment of categories to colors, is to add novel colors; e.g., yellow indicates X and blue indicates Y. Unfortunately, if novel features are added to the initially relevant dimension, it might be the case that differences in learnability of the dimensions were caused by differences in novelty. If novel features are added to both dimensions, it might be the case that differences in learnability are attributable to differences in degree of novelty, or differences in similarity of the novel values to the previous values, and so forth (Slamecka, 1968). This new design solves these problems by making the initial

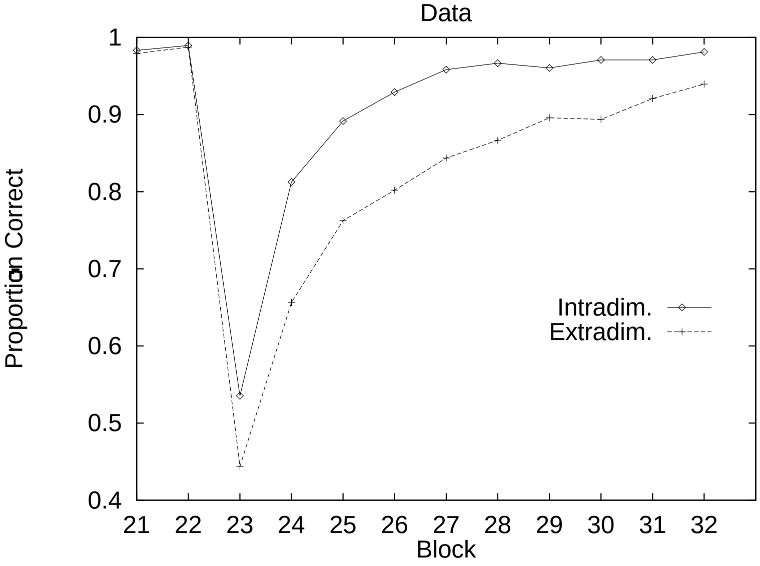

Figure 3. Results of the relevance shift experiment (adapted from Kruschke, 1996b).

problem involve *two* relevant dimensions, and no novel values at all in the shift phase.[2]

Human learning performance in this experiment is shown in Figure 3. It can be seen that people learned the intradimensional shift much faster than the extradimensional shift ($t(118) = 3.65$, $SE_{diff} = .026$, $p < .0001$ two-tailed).

Notice that the advantage of the intradimensional shift over the extradimensional shift cannot be explained by the number of exemplars that changed their route, because in both shift types there were four exemplars that changed their route. Another possible explanation for the difference is that only one dimension changed its relevance in the intradimensional shift, but all three dimensions changed their relevance in the extradimensional shift. This explanation is contradicted by results from another condition in the experiment (not summarized here) in which only two dimensions changed their relevance but the learning was even more difficult than the extradimensional shift.

This advantage of intradimensional over extradimensional shift has been found in many previous studies in many other species, but the results here are particularly compelling because the design involved no confounded variation of novelty. This robust difference should be addressed by any model of learning that purports to reflect learning by natural intelligent organisms.

A connectionist model with attentional learning

The advantage of intradimensional shift over extradimensional shift suggests that there is learned attention to dimensions. A model of learning should implement this explanatory principle.

[2] The original design used by Kruschke (1996b) also included two other types of shift: a complete reversal of all categories, and a change to another XOR structure with the previously irrelevant dimensions relevant and one previously irrelevant dimension relevant. These additional types of shift were useful for testing other hypotheses about shift learning.

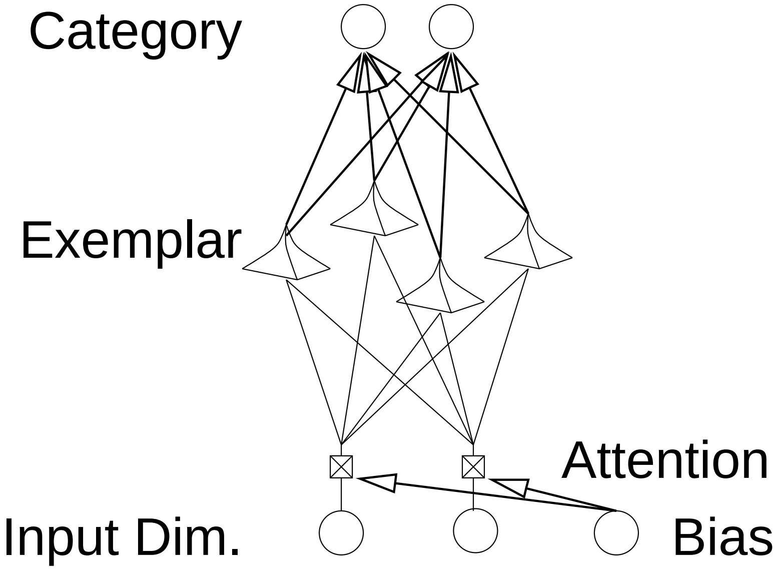

Figure 4. Architecture of model used for predictions of relevance shift experiments. Thicker arrows denote learned associative weights. The X's in boxes above the input dimensions represent the multiplicative weighting of the attention on the dimensions.

Any model of the relevance-shift experiment will also need to be able to learn the XOR category structure in the first phase of training. This structure is non-linear in the two relevant dimensions, meaning that no simple additive combination of the two relevant dimensions can accurately compute the correct categories. Instead, conjunctive combinations of dimensional values must be encoded in the model. There has been much research that suggests that people can and do encode configurations of values, also called *exemplars*, during learning (e.g., Nosofsky, 1992). The model to be fit to the shift-learning data formalizes this notion of exemplar representation, along with the notion of learned attention to dimensions.

The model fit to these data was called AMBRY by Kruschke (1996b) because it is a variant of the ALCOVE model (Kruschke, 1992). The architecture of (part of) AMBRY is shown in Figure 4. All aspects of the model have specific psychological motivations, and formalize explicit explanatory principles. Because of this correspondence between model parts and explanatory principles, the principles can be tested for their importance by excising the corresponding aspect of the model. In particular, the attentional mechanism can be functionally removed, and the restricted model can be tested for its ability to fit to data.

Activation propagation. In AMBRY, each dimension is encoded by a separate input node. If ψ_i denotes the psychological scale value of the stimulus on dimension i, then the activation of input node i is simply that scale value:

$$a_i^{\text{in}} = \psi_i. \tag{1}$$

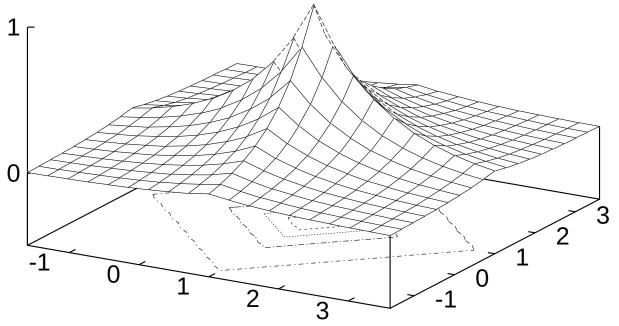

Figure 5. Activation function of an exemplar node. The surface shows a_j^{ex} of Equation 2 for a two-dimensional stimulus space, with $c = 1$, $\alpha_1 = 1$, $\alpha_2 = 1$, $\psi_{j1} = 1$ and $\psi_{j2} = 1$. The diamonds on the plane underneath the surface indicate the level contours of the surface.

Because the experiment counter-balanced the assignment of physical dimensions in Figure 1 to abstract dimensions in Figure 2, the dimensional values were simply assumed to be 1.0 and 2.0; e.g., for the short car, $\psi_{\text{height}} = 1.0$ and for the tall car, $\psi_{\text{height}} = 2.0$.

There is one exemplar node established for each of the eight cars. The activation of an exemplar node corresponds to the psychological *similarity* of the current stimulus to the exemplar represented by the node. Similarity drops off exponentially with distance in psychological space, as argued by Shepard (1987), and distance is computed using a city-block metric for psychologically separable dimensions (Garner, 1974; Shepard, 1964). An exemplar node is significantly activated only by stimuli that are fairly similar to the exemplar represented by the node. In other words, each exemplar node has a limited "receptive field" in stimulus space. Formally, the activation value is given by

$$a_j^{\text{ex}} = \exp\left(-c \sum_i \alpha_i |\psi_{ji} - a_i^{\text{in}}| \right) \qquad (2)$$

where c is a constant called the *specificity* that determines the narrowness of the receptive field, where α_i is the *attention strength* on the i^{th} dimension, and where ψ_{ji} is the scale value of the j^{th} exemplar on the i^{th} dimension. Because stimulus values are either 1.0 or 2.0, the values of ψ_{ji} are either 1.0 or 2.0. Figure 5 shows the activation profile of an exemplar node in a two-dimensional stimulus space. It is this pyramid-shaped activation profile that is used to represent the exemplar nodes in Figure 4.

Importantly, Equation 2 implies that increasing the attention strength on a dimension has the effect of magnifying differences on that dimension, so that differences along the dimension have a larger influence on the similarity. Thus, if a dimension is relevant to a categorization, the attention strength on that dimension can be increased to better distinguish the exemplars from the two categories. On the other hand, an irrelevant dimension can have its attention decreased, so that differences along that dimension do not needlessly impede learning. As will be explained below, AMBRY *learns* how to adjust the dimensional attention strengths to facilitate categorization. The attention strengths are indicated in Figure 4 by the arrows from a "bias" node (which is always activated) to the boxes marked with X's above the input nodes. The boxes are marked with X's to indicate that each attentional strength is a multiplier on the input.

Activation from the exemplar nodes is propagated to category nodes via weighted connections, illustrated in Figure 4 by the arrows from exemplar nodes to category nodes. The activation of each category node is determined by a standard linear combination of weighted exemplar-node activations. Finally, the activations of the category nodes are converted to choice probabilities by a ratio rule, such that the probability of choosing a category corresponds with the activation of the category relative to the total activation of all categories. The mathematical details of these operations are not critical for the present discussion, and can be found in the original article (Kruschke, 1996b).

Learning of attention and associations. The association weights between the exemplar nodes and the category nodes are learned by standard backpropagation of error (Rumelhart et al., 1986). Just as human learners are told the correct answer on each trial, the model is told the desired activation of the category nodes on each trial. Any discrepancy between the desired activation and the model-generated activation constitutes an error. The model then adjusts the dimensional attention values and the associative weights in such a way that the error is reduced as quickly as possible. Not necessarily all the error is eliminated on a single trial. This type of error reduction is called "gradient descent" because it is based on computing the derivative of the error with respect to the attention strengths or associative weights. Formulas for these derivatives are provided by Kruschke (1996b).

Of importance here is to note that the learning of the attention strengths is based on error reduction, and the amount or speed of learning is governed by a single parameter called the attentional learning rate. When this attentional learning rate is fixed at zero, the model has no ability to learn to selectively attend to relevant dimensions (but it can still learn categorizations because of the learnable association weights between exemplars and categories). By testing whether the model can fit the empirical data with its attentional learning rate set to zero, we can discover whether attentional learning is an essential principle in the model.

This model is an algorithmic description of learning. The model makes no claims about physical implementation. Different species might neurally implement the algorithm, or approximations to the algorithm, in different ways. In particular, there is no claim that nodes in the model correspond to neurons in the brain, nor is there any claim that gradient descent on error gets implemented as backpropagation of error signals through neural synapses. The model is therefore referred to as a type of *connectionist* model, and is never referred to as a *neural network* model.

Fit of the model. The top graph of Figure 6 shows the predictions of AMBRY when fit to the data shown in Figure 3. (In fact, these are the predictions when AMBRY is simultaneously fit to two other shift conditions, not discussed here. If AMBRY were fit only to the intra- and extra-

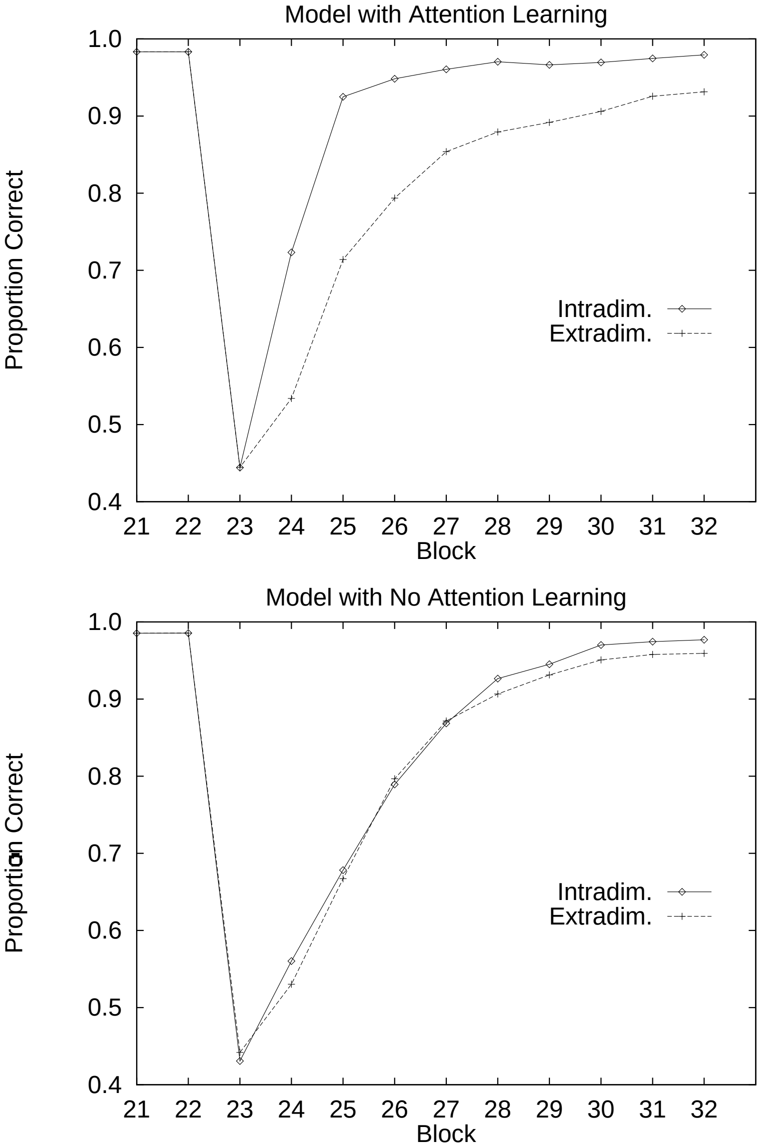

Figure 6. Model predictions for the relevance shift experiment. Top panel shows fit of model with attentional learning. Bottom panel shows fit of restricted model with no attentional learning. (Adapted from Kruschke, 1996b, .)

dimensional shifts, the predictions would be even closer to the data.) AMBRY shows the advantage of intradimensional shifts very robustly. It accommodates the data by learning in the first phase to attend to the two relevant dimensions and to ignore the irrelevant dimension. This learned attentional distribution must be unlearned in the extradimensional shift, and therefore the extradimensional shift is more difficult for the model.

Attentional learning is critical to account for the data. When the attentional learning rate is fixed at zero, the best fit of the model exhibits no difference between the types of shift, as can be seen in the bottom graph of Figure 6. The best the model can do without attentional learning is settle on the mean of the two types of shift.

Blocking of associative learning

Suppose two cues, A and B, are presented to a learner, followed by an outcome. Typically both cues will acquire moderate associative strength with the outcome. On the other hand, if the subject was trained in a previous phase to learn that A by itself predicts the outcome, then the associative strength from B seems to be very weak. It appears that the prior training with A has blocked, i.e. prevented, learning about B, despite the fact that B is now just as predictive of the outcome as A is. The phenomenon of blocking, first reported by Kamin (1968), is ubiquitous, occurring in many different procedural paradigms and in many different species.

“No empirical finding in the study of animal learning has been of greater theoretical importance than the phenomenon of blocking” (Williams, 1999, p. 618). Blocking is important because it shows clearly that associative learning is not based on merely the co-occurrence of cue and outcome. This fact contradicts a whole raft of learning models that increment associative strength whenever a cue and outcome co-occur.

For more than thirty years there have been two prominent theories of blocking. The dominant theory, formalized in the Rescorla-Wagner model (1972) and equivalent to the delta-rule of connectionist models, argues that learning is error-driven. Because the subject has already learned that cue A predicts the outcome, when cue B occurs there is no error in prediction and hence no learning. The Rescorla-Wagner model was motivated to a large degree by the phenomenon of blocking, and the model has been monumentally influential (Miller, Barnet, & Grahame, 1995; Siegel & Allan, 1996).

A competing theory, first suggested by Sutherland & Mackintosh (1971) and extended by Mackintosh (1975), claims that there is in fact something learned about the redundant relevant cue; namely, that it is irrelevant. In other words, subjects learn to suppress attention to the redundant cue. As was emphasized in the previous section regarding intra- and interdimensional shifts, learned attention can be assessed by measuring the difficulty of subsequent learning. Mackintosh & Turner (1971) measured how quickly a previously blocked cue could be learned about by rats, and found evidence in favor of the learned attention theory. Kruschke & Blair (2000) extended and expanded their experimental design in a study with humans, and found robust evidence that people learn to suppress attention to a blocked cue. Learning about a blocked cue was much weaker than learning about a non-blocked control cue.

Table 1: Design of experiment assessing discrimination learning after blocking.

Training I	A→1	E→4
Training II	A.B→1	E.F→4
(Blocking of B, F & G)	A.B→1	E.G→4
Training III	B.C→2	H.F→5
(Discrimination	B.D→3	H.G→6
learning)	A→1	E→4
Testing	B.C, B.D	H.F, H.G
	A	E
	C.F, C.G, D.F, D.G	

Note. Letters denote symptoms, numerals denote diseases.

This ubiquitous learning phenomenon, blocking, involves learned attention. Models of natural learning should incorporate mechanisms of learned attention. The remainder of this section reports previously unpublished results demonstrating the effects of blocking on subsequent learning, and a connectionist model that uses learned attention. As in the previous section, it will be shown that when attentional learning in the model is "turned off," the model cannot exhibit the critical effects.

Experiment design and results

In an experiment conducted in my lab, people had to learn which symptoms indicated certain fictitious diseases. A learning trial might consist of the following sequence of events. First, a list of symptoms is presented on a computer screen, e.g., "back pain" and "blurred vision." The subject would then indicate which disease she or he thought was the correct diagnosis, by pressing a corresponding key. Then the correct response was displayed on the screen. In the initial trials the person would just be guessing, but after several trials she or he could learn the correct diagnoses.

Table 1 shows the design of this experiment. Symptoms are indicated by letters, and diseases are indicated by numerals. I will first describe the third phase of training. The central aspect of the third phase is that people must learn to discriminate diseases that share a symptom. Thus, disease 2 is indicated by symptoms B and C (denoted B.C→2), and disease 3 is indicated by symptoms B and D (denoted B.D→3). The two diseases share symptom B, and therefore learning the diseases might be somewhat difficult. The same structure is present for diseases 5 and 6: H.F→5 and H.G→6.

If attention to the shared symptom were suppressed, then discrimination learning should be easier. On the other hand, if attention to the distinctive symptoms were suppressed, then discrimination learning should be harder. The first two phases of training were designed to bring about just such suppression of attention. Notice that in the first phase, symptom A always indicates disease 1, which is denoted A→1. In the second phase, symptom B is paired with symptom A as a redundant relevant cue; i.e., AB→1. Hence symptom B should by blocked, and should suffer suppressed attention. Hence subsequent learning of B.C→2 and B.D→3 should be enhanced. By contrast, the first

Table 2: Choice percentages from the test phase of discrimination learning after blocking.

	Response Choice																	
	A			E			C/D			F/G			C/Do			F/Go		
Symptoms	H	M	R	H	M	R	H	M	R	H	M	R	H	M	R	H	M	R
A	94	95	94	3	1	1	2	1	1	1	1	1	1	1	1	1	1	1
E	3	1	1	95	95	94	1	1	1	1	1	1	0	1	1	1	1	1
BC/BD	1	5	4	1	5	4	74	76	75	5	5	4	10	6	9	9	5	4
HF/HG	1	4	4	1	4	4	7	4	4	71	74	75	7	4	4	13	9	9
CF/CG/DF/DG	2	5	5	2	5	5	**49**	**46**	**41**	**33**	**34**	**41**	7	5	5	6	5	5

Note. Letters in left column denote symptom combinations summarized by corresponding row. Letters at top of columns denote the symptom corresponding to the disease selected. For example, the response choice "A" means the disease corresponding to symptom A, i.e., disease 1. The response choice "C/D" means the disease corresponding to symptom C if a test case including symptom C was presented, and the disease corresponding to symptom D if a test case including symptom D was presented. Under each choice, the column headed "H" indicates the human choice percentage, the column headed "M" indicates the full model percentage, and the column headed "R" indicates the restricted model percentage with no attention learning.

two phases are designed to block the distinctive symptoms F and G, so that learning of H.F→5 and H.G→6 should be worsened.[3]

The final testing phase is an additional assessment of the relative strengths of association established for the diseases that had a blocked shared symptom versus the diseases that had blocked distinctive symptoms. The test cases C.F, C.G, D.F and D.G present conflicting symptoms, so that their relative strengths can be directly assessed. In these cases, people should select the diseases associated with C and D more than the diseases associated with F and G.

There were 40 trials of training phase I, 80 trials of phase II, and 60 trials of Phase III, followed by 20 test trials. There was no test of blocking after phase 2 because numerous experiments in my lab have shown very robust blocking in this type of procedure (e.g. Kruschke & Blair, 2000). We can safely assume that blocking occurred. The eight symptoms were randomly selected for each subject from the following nine: ear ache, skin rash, back pain, dizziness, nausea, insomnia, bad breath, blurred vision, and nose bleed. Response keys (disease labels) were D, F, G, H, J, and K, randomly assigned to diseases for each subject.

A total of 89 students volunteered to participate for partial credit in an introductory psychology course. The results confirmed the predictions of attentional learning. In first third of training phase 3, people had higher accuracy on the diseases with a blocked shared symptom (46.3% correct) than on the diseases with the blocked distinctive symptoms (40.2% correct), t(88)=2.97, p=.004. (Collapsed across all of the third phase, the mean difference in percent correct was 2.9%.)

[3] Strictly speaking, the design does not constitute blocking of F and G, because these two symptoms are not perfectly correlated with the disease in phase 2. (Symptom B, however, does conform strictly to a blocking design.) Despite this departure from a strict blocking design, the theoretical implications regarding attentional learning remain the same.

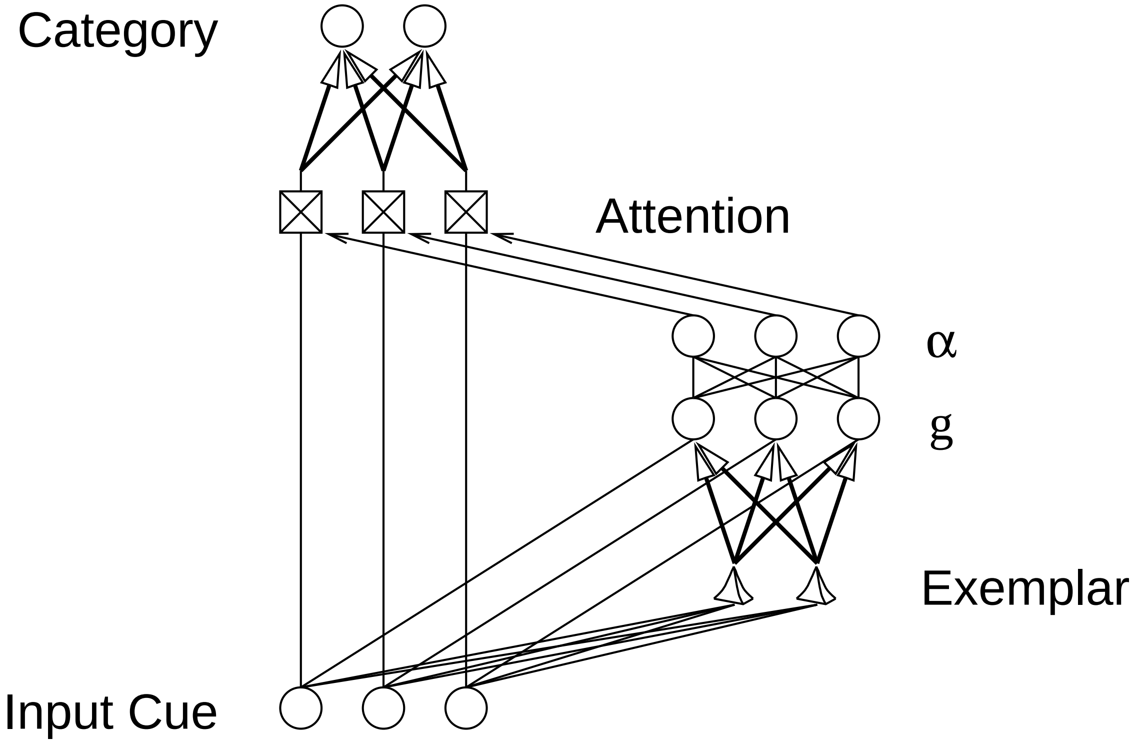

Figure 7. Architecture of the EXIT model. The thicker arrows denote learnable associative weights. The X's in boxes above the input cues represent the multiplicative weighting of the attention on the cues.

The final testing phase showed more robust effects of blocking. Table 2 shows the choice proportions for the five types of symptom combinations used in the test phase. Of most interest are the conflicting symptom pairs, C.F, C.G, D.F and D.G. In each of these cases a distinctive symptom (C or D) from the diseases with a blocked shared symptom is paired with a distinctive symptom (F or G) from the diseases with blocked distinctive symptoms. Attentional theory suggests that learning about symptoms C and D should be faster, hence stronger, than learning about symptoms F and G. Therefore in these cases of conflict the choices should favor the diseases corresponding with C and D over the diseases corresponding with F and G. The last row of Table 2 shows that this result did indeed occur, with 49% of the choices being for the C/D diseases and only 33% of the choices being for the F/G diseases. This difference is highly significant by a binomial test, $z = 4.62, p < .001$.

These results, along with those of Kruschke & Blair (2000), show that there is learned attention in blocking (at least in this type of procedure). Because blocking is so pervasive in natural learning, models of learning should address these attentional effects.

A connectionist model with attentional learning

Figure 7 depicts the ADIT model introduced by Kruschke (1996a) and extended by Kruschke (1999), referred to here as the EXIT model. EXIT is very similar in spirit to the AMBRY model described in the previous section (Figure 4). One difference between the models is that AMBRY used exemplar nodes between the inputs and the categories, whereas EXIT does not. This is merely a pragmatic simplification to reduce the number of free parameters, and is not a theoretical commitment. On the other hand, EXIT does have exemplars between the inputs and the attention nodes. The motivation for this is the idea that learned attentional distributions should be exemplar specific. For example, when a mushroom is smooth and flat, attention should be shifted to the flat shape, but attention should not *always* be shifted away from texture to shape.

Another difference between AMBRY and EXIT is that EXIT imposes a capacity constraint on the attention strengths. This capacity constraint in the model is supposed to reflect attentional capacity constraints in humans and other animals: If attention to a feature is increased, it must necessarily be decreased on another feature. This constraint is indicated in Figure 7 by the criss-crossing lines between the *gain* nodes and the attention nodes. When a feature is present, the gain node is activated by default, but can also be influenced by learned associations from exemplars. The gains on each feature are then normalized to produce the overall attention to each feature.

One last but important enhancement of EXIT is that attention shifts are executed rapidly within a trial, although they may be learned only gradually across trials. Thus, when corrective feedback is provided on a trial, the error drives a relatively large shift in attention before any associative weight changes are made. This shifted distribution of attention acts as the target values to be learned by the associative weights between the exemplars and the gain nodes. This gives EXIT one more free parameter: An attentional shift rate, distinct from the attentional learning rate.

Complete mathematical details of the model are provided elsewhere (Kruschke, 1999), but it might be useful here to describe the influence of attention in the formulas for categorization and associative weight change. These two formulas show clearly the separate roles of attention for response generation and for learning. Responses are generated proportionally to the activation of the category nodes, and these nodes are activated according to the summed weighted activation of the input cues. Formally, the activation of the k^{th} category node is determined as

$$a_k^{\text{cat}} = \sum_i w_{ki} \alpha_i a_i^{\text{in}} \tag{3}$$

where w_{ki} is the associative weight from the i^{th} cue to the k^{th} category, α_i is the attention allocated to the i^{th} cue, and a_i^{in} is the activation of the i^{th} cue. Notice that a cue has an influence on the category choice only to the extent that the cue is attended to. Associative weight changes are also affected by attention. The change in the associative weight from the i^{th} cue to the k^{th} output (denoted Δw_{ki}) is given by

$$\Delta w_{ki} = \left(t_k - a_k^{\text{cat}}\right) \alpha_i a_i^{\text{in}} \tag{4}$$

where t_k is the *teacher* value (correct response) for the k^{th} category node. Notice that a weight from a cue is changed only to the extent that the cue is being attended to. Thus, the model only learns about what is being attended to, and the model attends to whatever reduces error best.

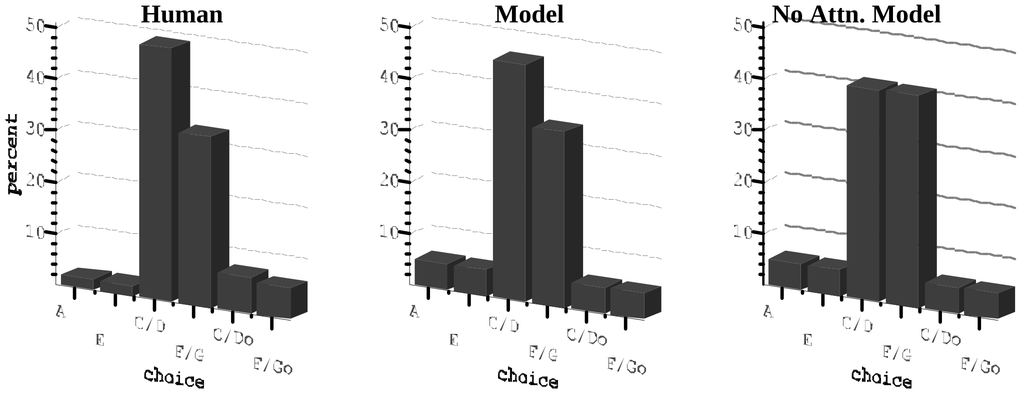

Figure 8. Results of the conflicting symptom tests of the blocking experiment. Data from last row of Table 2.

In summary, processing in the EXIT model occurs as follows: Cues are presented and activate the corresponding input nodes. By default, the corresponding attention gain nodes are then activated, modulated by any learned redistribution of attention for exemplars similar to the stimulus. The normalized (capacity constrained) attention then multiplicatively gates the cue activations propagated to the category nodes. Category node activations are mapped to response probabilities as in AMBRY. When the correct classification is provided, the error first drives a relatively large attention shift. After this shift is completed, the associative weights to the attentional gain nodes are adjusted to try to learn this new distribution of attention, and the associative weights to the category nodes are adjusted try to diminish any remaining predictive error.

Fit of the model. The model was fit to the testing phase data of Table 2 (and not to the third phase learning data) with each test type weighted by the number of distinct cases contributing to the type. The best fitting predictions of the model mirror the data quite well. In particular, EXIT shows a strong preference for C/D diseases over F/G diseases in the conflicting symptom cases (see the last row of Table 2). Although not fit to the third learning phase, EXIT, like humans, shows a small (2.2%) advantage for the diseases that had their shared symptom blocked.

Figure 8 displays data from the conflicting symptom tests (CF/CG/DF/DG) of Table 2 in the form of a bar graph. The left graph displays the human choice percentages; the middle graph displays the predictions of EXIT. In the human data, the two bars for the C/D and F/G choices are at distinctly different heights, as they are in the predictions of EXIT.

Importantly, when the attentional learning rate of EXIT is fixed at zero (but attentional shifting is still allowed), the critical effects cannot be produced by the model. The best fitting predictions of this “no attention” restricted version of the model are also shown in Table 2 and in the right graph of

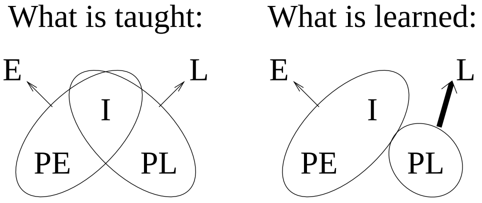

Figure 9. Structure of categories in the inverse base rate effect.

Figure 8. Notice that there is *no* difference between C/D and F/G response proportions for conflict tests (last row of table). Moreover, there is *zero* difference between groups in the third learning phase when attention learning was disallowed. Thus, attention learning is crucial for the model to exhibit the effects observed in people.

The inverse base rate effect

The learning behavior reviewed above can be thought of as irrational and suboptimal. Consider the advantage of intradimensional over extradimensional shift. The two structures in the second phase of learning (see Figure 2) are isomorphic, and so one might think that an optimal learning device should learn them equally efficiently. Consider the phenomenon of blocking. The redundant relevant cue in phase 2 (see Table 1) is, after all, perfectly predictive of the outcome, and so a rational and optimal learner should learn about the cue. Moreover, the two pairs of diseases that share a symptom in phase 3 are isomorphic, and so they should be learned equally well.

Another prominent example of apparently irrational learning is the inverse base rate effect. Suppose that very frequently symptoms I and PE occur together and always indicate disease E (I.PE→E), as shown in the left panel of Figure 9. On some rare occasions, symptoms I and PL occur, and when they do, they always indicate disease L (I.PL→L). Notice that symptom I is shared by both diseases, and therefore is an imperfect predictor of the diseases. Symptom PE is a perfect predictor of disease E, and symptom PL is a perfect predictor of disease L. The left panel of Figure 9 shows this structural symmetry.

After learning these diseases, experiment participants are asked to make diagnoses for novel combinations of symptoms, such as PE.PL and I by itself. For symptom I by itself, which objectively is equally indicative of both diseases, people tend to choose disease E, which is appropriate because disease E has a larger base rate; i.e., a higher frequency of occurrence. For symptoms PE.PL, however, people strongly tend to choose the rare disease L. This pattern of results was dubbed the "inverse base rate effect" by Medin & Edelson (1988). This effect is found in disease diagnosis procedures (Kruschke, 1996a; Medin & Edelson, 1988), in a random-word association procedure (Dennis & Kruschke, 1998), and in a geometric-figure association procedure (Fagot, Kruschke, Dépy, & Vauclair, 1998), so it is a very robust phenomenon. It has not yet been reported in other species, however.

Kruschke (1996a) explained the inverse base rate effect as a consequence of rapidly shifting attention. One strong consequence of different base rates is that people tend to learn about the fre-

Table 3: Design of experiment assessing learned attention in the phased inverse base rate effect.

Training I	I1.PE1→E1	I2.PE2→E2
Training II	I1.PE1→E1	I2.PE2→E2
	I1.PL1→L1	I2.PL2→L2
Testing	I, PE.PL, etc.	
Training III	I1.PE1→N1	I2.PE1→N2
(easy, I in PE)	I1.PE2→N1	I2.PE2→N2
Training III	I1.PL1→N1	I2.PL1→N2
(hard, I in PL)	I1.PL2→N1	I2.PL2→N2

Note. E, L and N denote disease. I, PE, PL denote symptoms.

quent disease before they learn about the rare disease. Thus, people learn early that symptoms I and PE each have a moderate associative strength with disease E. Then people learn later about cases of the rare disease, I.PL→L. As argued in the introductory discussion of learning about mushrooms, people shift attention away from the symptom I already associated with the common disease E, and learn predominantly about the distinctive symptom of the rare disease, thereby building up a strong association from symptom PL to disease L. This asymmetry in the learned associations about the diseases is illustrated in the right panel of Figure 9. Consequently, when tested with PE.PL, the strong association from PL to L dominates the moderate association from PE to E. Further empirical and modeling results added supportive evidence to this explanation in terms of rapidly shifting attention.

The emphasis of Kruschke (1996a) was rapidly shifting attention during single trials of learning, rather than on learned redistributions of attention. Nevertheless, the notion that attentional redistributions are learned suggests that additional learning *subsequent* to the inverse base rate effect should be impacted by learned attention shifts. In particular, for symptom pair I.PL, attention should be shifted away from I to PL, so that subsequent learning about I in the context of PL should be difficult. On the other hand, for symptom pair I.PE, attention is not strongly shifted away from I, so that subsequent learning about I in the context of PE should be relatively easy. Results of an experiment that tests this prediction are presented here. It will be shown that the results can be fit by EXIT with attentional shifting, but when attentional shifting is “turned off,” the restricted model fails.

Experiment design and results

Table 3 shows the design of an experiment that assesses learned attention in a phased-training version of the inverse base rate effect. The first two phases force people to learn I.PE→E before learning I.PL→L, instead of relying on base rates to accomplish indirectly this ordering of learning. The design incorporates two copies of the same basic structure, so, for example, the first phase consists of I1.PE1→E1 and I2.PE2→E2. After the second training phase, there is a test phase to measure the magnitude of the inverse base rate effect. This much of the design is a replication of previous studies (e.g., Kruschke, 1996a, Exp. 2).

Table 4: Results from test phase of experiment assessing discrimination learning after blocking, with prediction of model in parentheses.

	Response Choice											
	E			L			Eo			Lo		
Symptoms	H	M	R	H	M	R	H	M	R	H	M	R
I.PE	93	94	93	3	2	3	2	2	2	1	2	2
I.PL	12	9	3	85	85	93	0	3	2	3	3	2
I	**75**	**84**	**44**	**17**	**5**	**45**	5	5	6	3	5	6
PE.PL	**28**	**24**	**39**	**67**	**68**	**39**	1	4	11	4	4	11
PE.PLo	**32**	**24**	**50**	6	4	1	3	4	1	**59**	**68**	**48**
I.PE.PL	52	48	43	43	46	43	1	3	7	4	3	7
I.PE.PLo	63	61	67	5	3	5	3	3	2	29	33	26

Note. Results are collapsed across pairs 1 and 2. For example, Symptom I refers to cases of I1 and I2. If I1 was presented, then Choices E, L, Eo and Lo refer to diseases E1, L1, E2 and L2, respectively. If I2 was presented, then Choices E, L, Eo and Lo refer to diseases E2, L2, E1 and L1, respectively. PE.PLo indicates cases of PE1.PL2 and PE2.PL1 combined. Under each choice, the column headed "H" indicates the human choice percentage, the column headed "M" indicates the full model percentage, and the column headed "R" indicates the restricted model percentage with no attention learning.

The third phase of training breaks new ground. Half the subjects went on in this phase to learn that symptoms I1 and I2 were relevant to diagnosing new disease N1 and N2, in the context of PE1 and PE2. This was predicted to be relatively easy. The other half of the subjects learned that I1 and I2 were relevant in the context of PL1 and PL2. This was predicted to be relatively difficult.

There were 40 trials of training phase I, 80 trials of phase II, 28 trials of the testing phase, and 80 trials of Phase III. Symptoms and response keys were selected as in the blocking experiment.

A total of 83 students volunteered to participate for partial credit in an introductory psychology course. Six subjects' data were excluded from further analysis because they failed to reach 80% correct in the last half of the second phase of training. This left 38 subjects in the "easy" condition and 39 in the "hard" condition. There were no differences between groups in the first two phases of learning. Table 4 shows the choice percentages in the test phase, collapsed across the two groups. For test case I, choices for E were far greater than choice for L (75% vs. 17%), $\chi^2(1,245)/4 = 25.15$, $p < .001$. For test case PE.PL, choices for L were far greater than choices for E (67% vs. 28%), $\chi^2(1,262)/4 = 11.13$, $p < .001$. For test case PE.PLo, choices for Lo were far greater than choices for E (59% vs. 32%), $\chi^2(1,225)/4 = 5.60$, $p < .025$. Thus the inverse base rate effect is strongly in evidence.

The main novel result regards the relative ease of learning in the third phase. Collapsed across all blocks of the third phase, the mean percents correct were 82.7% for the "easy" group and 76.2%

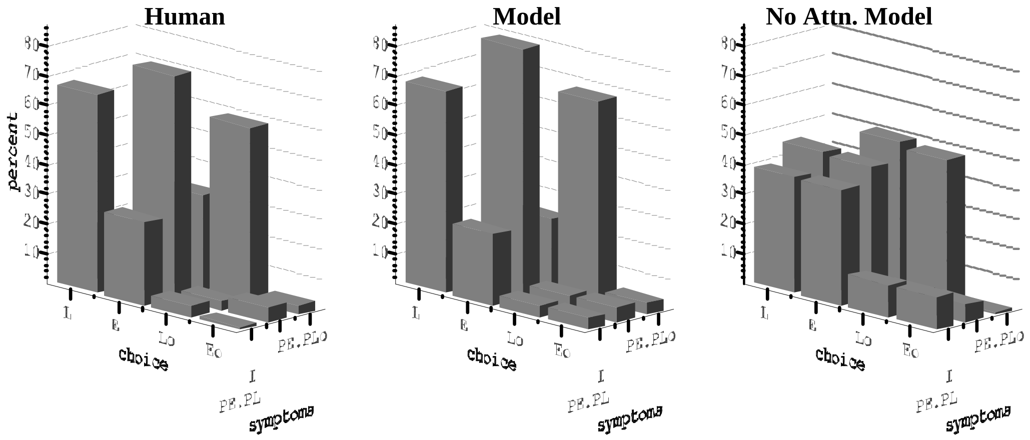

Figure 10. Selected data from inverse base rate experiment, with predictions of EXIT model with or without attentional shifting and learning.

for the “hard” group. These are reliably different, $t(75) = 2.10$, $SE_{diff} = .031$, $p = .040$ two-tailed for unequal-variance corrected df of 62.10. This difference cannot be attributed to group differences in the initial phases, because there were no hints of differences between groups in the first two phases.

Fit of the model. The model, EXIT, was fit to the testing data and to the third phase training data. Table 4 shows that the model predictions fit the testing data well, with the model exhibiting a robust preference for E when given symptom I alone, a strong preference for L when presented with symptom pair PE.PL, and a strong preference for Lo when presented with PE.PLo. The model also fit the learning in the third phase well, predicting 80.9% correct in the easy condition and 72.5% correct in the hard condition, compared with 82.6% and 76.2% by humans.

Selected (bold font) data from Table 4 are presented graphically in Figure 10. It can be seen from the left and middle graphs of Figure 10 that the model reproduces the pattern of results seen in the human data quite well.

When the attention shifting was turned off, so that the shift rate had a fixed value of zero (and hence there was no attentional learning, either), the restricted model was entirely unable to show the trends of interest. Table 4 shows that the model without attention exhibits *no* preference for E when given symptom I alone, *no* preference for L when presented with symptom pair PE.PL, and *no* preference for Lo when presented with symptom pair PE.PLo. The model also shows *no* difference between groups in the third phase of learning, predicting 78.2% correct in the easy condition and 78.3% correct in the hard condition. Yet again we see that attentional mechanisms are critical for the model to fit the data.

The right graph of Figure 10 shows clearly the failure of the restricted model to capture the preferences shown by humans. Without attention shifting and learning, the model can learn the training patterns, but generalizes nothing like people do, neither in the test phase nor in the subsequent training phase.

Summary and conclusion

The preceding sections provided three illustrations of the importance of attentional shifts and attentional learning in models of natural learning. In all three experiments, attentional learning in one phase was assessed by examining the ease of learning in a subsequent phase. If people have learned to attend to one feature or ignore another feature, then subsequent learning about the attended-to feature should be easier than learning about the ignored feature. In the first illustration, an advantage of intradimensional shift over extradimensional shift was demonstrated with an experiment (Kruschke, 1996b) that avoided a problem common to all previous designs, i.e., confounded changes in novelty. The second illustration showed that the pervasive phenomenon of blocking involves learned suppression of attention to the blocked cue. The experiment demonstrated that discrimination learning was easier when the shared cue was previously blocked than when the distinctive cues were previously blocked. The third illustration suggested that the rapid attention shifting evident in the inverse base rate effect also involves learned attention shifts, because subsequent learning about the imperfect predictor was more difficult in the context of the later learned (attended to) distinctive cue than in the context of the earlier learned (less strongly attended to) distinctive cue.

All three phenomena —intradimensional shift advantage, blocking, and the inverse base rate effect— have been found in a variety of procedural paradigms and settings. The first two have been found in a variety of animal species (and the third is relatively recent and has not yet been systematically sought in other species). Therefore attentional learning is a widespread phenomenon and should not be ignored by those who wish to model natural learning.

In all three illustrations, connectionist models that directly implemented attentional shifting and learning fit the data nicely. When the attentional shifting and learning was "turned off," the models could not exhibit the signature effects observed in the human data. The modeling adds supportive evidence to the veracity of the attentional theory.

Relation to other learning models

There have been a variety of connectionist models of associative learning proposed in recent years, some of which incorporate notions of attention, yet none of which address the type of attentional phenomena described here.

Gluck & Bower (1988) proposed a simple linear associator, that learned by the delta rule, as a model of apparent base rate neglect in human learning. Their seminal article initiated a series of further investigations by several researchers, demonstrating the robustness of the empirical effect and of the model's ability to address it. Yet it turned out that their model cannot account for results from a modest parametric variation of their experimental design, and instead an enhanced model that incorporates rapidly shifting attention is sufficient (Kruschke, 1996a).

Shanks (1992) proposed a variation of a linear associator, called the attentional connectionist model (ACM), in which the attention allocated to a cue is inversely related to the cue's base rate. Thus, the attention allocated to a cue corresponds to the cue's surprisingness or novelty. This notion of attention in the ACM is quite different than the notion expounded in this chapter. The attention in ACM does not shift rapidly in response to categorization errors. Kruschke (1996a) showed that the ACM does not fit data from an experiment examining apparent base rate neglect. Nevertheless, future empirical data might demand the inclusion of ACM-style novelty-based attention in addition to rapidly shifting error-driven attention.

Nosofsky, Gluck, Palmeri, McKinley, & Glauthier (1994) described a model that maintains separate learning rates for each feature and each combination of features. These individual learning rates, or associabilities, are adjusted in response to error. This interesting approach was first proposed by Sutton and colleagues (e.g., Gluck, Glauthier, & Sutton, 1992; Jacobs, 1988). The model was able to capture aspects of a classic learning study that the authors replicated, but the model did not fit the data quantitatively as well as ALCOVE. This approach is intriguing and deserves further investigation. It might be particularly challenged, however, by learning phenomena that are produced by *rapid* attention shifts or by exemplar-specific learned attention.

Some neurally-inspired models of learning, such as those of Schmajuk & DiCarlo (1992) and of Gluck & Myers (1997), implement types of attentional modulation in learning. In these models, however, the attentional modulation affects all cues simultaneously, and does not rapidly select component cues within an array. It may well turn out that both types of attentional mechanism are needed in a comprehensive model of learning.

The configural model of Pearce (1994) incorporates exemplar nodes similar to AMBRY, and it incorporates attentional normalization similar to EXIT, but it does not incorporate any kind of shifting selective attention. It therefore is unable to address the effects highlighted in this chapter.

The Rational model of categorization (Anderson, 1990) is motivated by normative calculation of conditional probabilities, such that the learner is assumed to be accumulating statistics about feature and category co-occurrences, and then classifying items according to their Bayesian probabilities. The rational model can be implemented in a network framework not unlike a connectionist model (Anderson, 1990, p. 137). The model can account for many findings in learning, but one of phenomenon it does not address is the inverse base rate effect. In particular, it fails to show an inverse base rate effect for test symptoms PE.PLo, whereas humans show a strong effect (see Table 4 and Figure 10). The Rational model has no mechanism for shifting or learning attention.

Mackintosh's (1975) classic model for attention learning was invented as a direct formal expression of intuitions about how attention works, based on empirical findings. The formalism was not couched in any larger-scale framework to explain what the model mechanism accomplished computationally, or how the attentional mechanism related to the associative weight learning mechanism. Connectionist modeling adds such a larger-scale perspective. The EXIT model described in this chapter has an architecture motivated by psychological principles similar to Mackintosh's. But the mechanisms for attention shifting, attention learning, and associative weight learning are all derived by a common goal: error reduction. It turns out that a special case of EXIT is very nearly identical to the formulas proposed by Mackintosh (1975) (see Kruschke, 1999).

Attentional shifting and learning are rational

Attention shifts and learned attention are good for rapid learning of new associations without damaging previously learned associations. While this accelerates learning, it can also lead to apparently irrational behaviors. The irrationality of intradimensional shift advantage, blocking, and the inverse base rate effect was described above. There are many other examples. Consider a situation wherein a cue is only imperfectly correlated with an outcome. The extent to which people (and other animals) learn to utilize the cue decreases when other, irrelevant cues are added (e.g., Castellan, 1973; Wagner, Logan, Haberlandt, & Price, 1968). For an optimal learner, the presence of irrelevant information should not affect the ultimate utilization of relevant information, yet for natural learners it does. Kruschke & Johansen (1999) review a number of related phenomena in probabilistic category learning, and address a panoply of irrational behavior with a connectionist model called RASHNL (which stands for Rapid Attention SHifting 'N' Learning). The rash shifts of attention facilitate the rational goal of rapid learning, but also lead to over- or under-commitments to various sources of information. Thus, a model that is driven purely by rapid error reduction can generate a number of seemingly irrational behaviors, just like people and many other animals. The pervasiveness of these learning phenomena, across situations and across species, suggests that it would be irrational for connectionist modelers to ignore attentional learning.

References

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

Anderson, J. R. (1990). *The Adaptive Character of Thought*. Hillsdale, NJ: Erlbaum.

Castellan, N. J. (1973). Multiple-cue probability learning with irrelevant cues. *Organizational Behavior and Human Performance, 9*, 16–29.

Dennis, S. & Kruschke, J. K. (1998). Shifting attention in cued recall. *Australian Journal of Psychology, 50*, 131–138.

Fagot, J., Kruschke, J. K., Dépy, D., & Vauclair, J. (1998). Associative learning in baboons (papio papio) and humans (homo sapiens): species differences in learned attention to visual features. *Animal Cognition*, *1*, 123–133.

Garner, W. R. (1974). *The Processing of Information and Structure*. Hillsdale, NJ: Erlbaum.

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

Gluck, M. A., Glauthier, P. T., & Sutton, R. S. (1992). Adaptation of cue-specific learning rates in network models of human category learning. In *Proceedings of the fourteenth annual conference of the Cognitive Science Society*, pp. 540–545. Hillsdale, NJ: Erlbaum.

Gluck, M. A. & Myers, C. E. (1997). Psychobiological models of hippocampal function in learning and memory. *Annual Review of Psychology, 48*, 481–514.

Hopfield, J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences USA, 79*.

Jacobs, R. A. (1988). Increased rates of convergence through learning rate adaptation. *Neural Networks, 1*, 295–307.

Kamin, L. J. (1968). 'Attention-like' processes in classical conditioning. In M. R. Jones (Ed.), *Miami symposium on the prediction of behavior: Aversive stimulation*, pp. 9–33. Coral Gables, FL: University of Miami Press.

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

Kruschke, J. K. (1996a). Base rates in category learning. *Journal of Experimental Psychology: Learning, Memory & Cognition*, *22*, 3–26.

Kruschke, J. K. (1996b). Dimensional relevance shifts in category learning. *Connection Science, 8*, 201–223.

Kruschke, J. K. (1999). Toward a unified model of attention in associative learning. Revision to appear in *The Journal of Mathematical Psychology*. Available online http:// www.indiana.edu/ ¯kruschke/ tumaal.html.

Kruschke, J. K. & Blair, N. J. (2000). Blocking and backward blocking involve learned inattention. *Psychonomic Bulletin & Review, 00*, 000–000. In press. Available from http:// www.indiana.edu/ ¯kruschke/ kb99.html.

Kruschke, J. K. & Johansen, M. K. (1999). A model of probabilistic category learning. *Journal of Experimental Psychology: Learning, Memory & Cognition*, *25*, 1083–1119.

Mackintosh, N. J. (1965). Selective attention in animal discrimination learning. *Psychological Bulletin, 64*, 124–150.

Mackintosh, N. J. (1975). A theory of attention: Variations in the associability of stimuli with reinforcement. *Psychological Review, 82*, 276–298.

Mackintosh, N. J. & Turner, C. (1971). Blocking as a function of novelty of CS and predictability of UCS. *Quarterly Journal of Experimental Psychology, 23*, 359–366.

Medin, D. L. & Edelson, S. M. (1988). Problem structure and the use of base-rate information from experience. *Journal of Experimental Psychology: General, 117*, 68–85.

Miller, R. R., Barnet, R. C., & Grahame, N. J. (1995). Assessment of the Rescorla-Wagner model. *Psychological Bulletin*, *117*, 363–386.

Minsky, M. L. & Papert, S. A. (1969). *Perceptrons*. Cambridge, MA: MIT Press. 1988 expanded edition.

Nosofsky, R. M. (1992). Exemplars, prototypes, and similarity rules. In A. F. Healy, S. M. Kosslyn, & R. M. Shiffrin (Eds.), *Essays in honor of William K. Estes. Vol. 2: From learning processes to cognitive processes*, pp. 149–167. Hillsdale, NJ: Erlbaum.

Nosofsky, R. M., Gluck, M. A., Palmeri, T. J., McKinley, S. C., & Glauthier, P. (1994). Comparing models of rule-based classification learning: A replication of Shepard, Hovland, and Jenkins (1961). *Memory & Cognition*, *22*, 352–369.

Pearce, J. M. (1994). Similarity and discrimination: A selective review and a connectionist model. *Psychological Review, 101*, 587–607.

Rescorla, R. A. & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and non-reinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical Conditioning: II. Current Research and Theory*, pp. 64–99. New York: Appleton-Century-Crofts.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review, 65*, 386–408.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In J. L. McClelland & D. E. Rumelhart (Eds.), *Parallel Distributed Processing,* Vol. 1, chap. 8, pp. 318–362. Cambridge, MA: MIT Press.

Schmajuk, N. A. & DiCarlo, J. J. (1992). Stimulus configuration, classical conditioning, and hippocampal function. *Psychological Review, 99*, 268–305.

Shanks, D. R. (1992). Connectionist accounts of the inverse base-rate effect in categorization. *Connection Science*, *4*, 3–18.

Shepard, R. N. (1964). Attention and the metric structure of the stimulus space. *Journal of Mathematical Psychology*, *1*, 54–87.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*, 1317–1323.

Siegel, S. & Allan, L. G. (1996). The widespread influence of the Rescorla-Wagner model. *Psychonomic Bulletin & Review*, *3*, 314–321.

Slamecka, N. J. (1968). A methodological analysis of shift paradigms in human discrimination learning. *Psychological Bulletin*, *69*, 423–438.

Sutherland, N. S. & Mackintosh, N. J. (1971). *Mechanisms of animal discrimination learning*. New York: Academic Press.

Wagner, A. R., Logan, F. A., Haberlandt, K., & Price, T. (1968). Stimulus selection in animal discrimination learning. *Journal of Experimental Psychology*, *76*, 171–180.

Williams, B. A. (1999). Associative competition in operant conditioning: Blocking the response-reinforcer association. *Psychonomic Bulletin & Review*, *6*, 618–623.

Wolff, J. L. (1967). Concept-shift and discrimination-reversal learning in humans. *Psychological Bulletin*, *68*, 369–408.