Version of November 12, 2009.
Kruschke, J. K. (2011). Models of attentional learning. In: E. M. Pothos and A. J. Wills (eds.), Formal Approaches in Categorization, pp. 120-152. Cambridge University Press.

Models of Attentional Learning

John K. Kruschke
Indiana University, Bloomington

Many theories of learning provide no role for selective attention (e.g., Anderson, 1991; Pearce, 1994; Rehder & Murphy, 2003). Selective attention is crucial, however, for explaining many phenomena in learning. The mechanism of selective attention in learning is also well motivated by its ability to minimize proactive interference and enhance generalization, thereby accelerating learning. Therefore, not only does the mechanism help explain behavioral phenomena, it makes sense that it should have evolved (Kruschke & Hullinger, 2010).

The phrase "learned selective attention" denotes three qualities. First, "attention" means the amplification or attenuation of the processing of stimuli. Second, "selective" refers to differentially amplifying and/or attenuating a subset of the components of the stimulus. This selectivity within a stimulus is different from attenuating or amplifying all aspects of a stimulus simultaneously (cf. Larrauri & Schmajuk, 2008). Third, "learned" denotes the idea that the allocation of selective processing is retained for future use. The allocation may be context sensitive, so that attention is allocated differently in different contexts.

There are many phenomena in human and animal learning that suggest the involvement of learned selective attention. The first part of this chapter briefly reviews some of those phenomena. The emphasis of the chapter is not the empirical phenomena, however. Instead, the focus is on a collection of models that formally express theories of learned attention. These models will be surveyed subsequently.

Phenomena suggestive of selective attention in learning

There are many phenomena in human and animal learning that suggest that learning involves allocating attention to informative cues, while ignoring uninformative cues. The following subsections indicate the benefits of selective allocation of attention, and illustrate the benefits with particular findings.

The author thanks Michael A. Erickson for discussion regarding ATRIUM and COVIS. Correspondence can be addressed to John K. Kruschke, Department of Psychological and Brain Sciences, Indiana University, 1101 E. 10th St., Bloomington IN 47405-7007, or via electronic mail to kruschke@indiana.edu. Supplementary information can be found at http://www.indiana.edu/~kruschke/

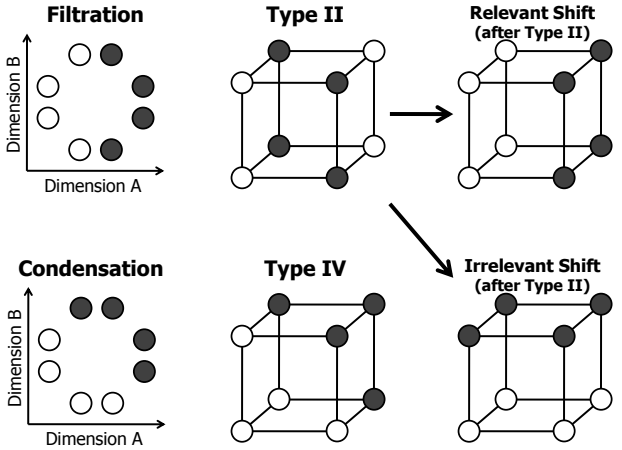

Figure 1. Category structures that illustrate benefits of selective attention. Axes denotes stimulus dimensions. Disks denote stimuli, with the color of the disk denoting the correct category label. The structures in the upper row are easier to learn than the corresponding structures in the lower row.

Attentional shifts facilitate learning

Learning can be faster when selective attention is able to enhance the relevant cues and suppress the irrelevant cues. As an example, consider the upper-left panel of Figure 1, labeled "Filtration". It shows a two-dimensional stimulus space, wherein each point represents a stimulus with corresponding values on the two dimensions. For example, Dimension B could be the height of a rectangle, and Dimension A could be the horizontal position of an interior line segment. The disks indicate the stimuli that were used on different training trials, and the color of the disk indicates the correct category label. Notice that for the Filtration structure, the correct category label can be inferred from Dimension A alone; Dimension B can be ignored. The lower-left panel shows a very similar structure labeled "Condensation". The only difference is in which stimuli belong to which category. Notice that for the Condensation structure, the correct category label can be inferred only by using information from both dimensions.

When people are trained on the Filtration and Condensation structures, the Filtration structure is learned much faster than the Condensation structure (Kruschke, 1993). This difference obtains despite the fact that both structures are linearly separable, and the clustering of instances is actually slightly better for the Condensation structure than for the Fil-

Table 1
Training designs for blocking and highlighting.

	Design			
Phase	Blocking		Highlighting	
Early	A→X	F→Y	I.PE→E	
Late	A.B→X	C.D→Y	I.PE→E	I.PL→L
Test	B.D→? (Y) A.C→? (X)		I→? (E) PE.PL→? (L)	

Note. Each cell indicates Cues→Correct Response. In the test phase, typical response tendencies are shown in parentheses.

tration structure. The Filtration advantage can be naturally explained by positing selective attention. In the Filtration structure, people learn to pay attention to the relevant dimension and they learn to ignore the irrelevant dimension. The selective attention enhances discriminability along the relevant dimension and greatly enhances generalization across values of the irrelevant dimension.

Another example is shown in the middle column of Figure 1. These structures involve stimuli with three binary dimensions; e.g., big/small, red/blue, triangle/square. Inspection of the upper-middle structure, labeled "Type II", reveals that the vertical dimension is irrelevant; i.e., it can be ignored without loss of accuracy. The remaining two dimensions are relevant, and the categorization is a non-linearly separable, exclusive-OR on those dimensions. The lower-middle structure, labeled "Type IV", merely re-arranges the category assignments of the exemplars. In this structure, no dimension can be ignored if perfect accuracy is to be attained. The structure is linearly separable, however, and comprises two prototypes in opposite corners of the stimulus space.

When people are trained on Type II and Type IV, Type II is faster to be learned, despite the fact that it involves a non-linearly separable categorization (Nosofsky, Gluck, Palmeri, McKinley, & Glauthier, 1994; Shepard, Hovland, & Jenkins, 1961). Again this difference can be explained by the action of selective attention. When learning Type II, people learn to ignore the irrelevant dimension, thereby quickly generalizing across that dimension.

Attentional shifting protects previous learning and accelerates new learning

Consider a situation in which a new response is to be learned. In this situation, the stimuli that occur for the new response have some new components but also some components that have been previously associated with old responses. The old components could cause proactive interference when trying to learn the new response. However, if the old components could be selectively suppressed when learning the new response, learning of the new response would be accelerated, while the previously learned association would be protected.

This sort of phenomenon has been observed in a training design referred to as "highlighting", and displayed in the right side of Table 1. The stimuli consist of selectively attendable cues, such as spatially separated words displayed on a computer screen. In the early phase of training, two cues, denoted I and PE, occur with outcome E. This case is denoted I.PE→E. In the late phase of training, those cases continue, interspersed with cases of a new correspondence: I.PL→L. The total number of trials of I.PE→E equals the total number of trials of I.PL→L; there is merely a front-loading of the I.PE→E cases to assure that they are learned first. Notice that the two cases are symmetric: The early outcome E is indicated by a perfect predictor PE and an imperfect predictor I, while the late outcome L is indicated by a perfect predictor PL and the imperfect predictor I. If people learn this simple symmetry, then cue I should be equally associated with outcomes E and L. In fact, people strongly prefer to respond with E when tested on cue I. This preference cannot be trivially explained as a mere primacy effect, however, because when people are presented with the cue pair PE.PL, people strongly prefer the outcome L (for a review, see Kruschke, 2010). This "torsion" in the response preferences, whereby E is preferred for one ambiguous cue but L is preferred for another ambiguous cue, is called the highlighting effect.

The highlighting effect can be explained by the action of selective attention during learning. During the early phase, people build moderate-strength associations from cues I and PE to outcome E. During the later phase, when learning I.PL→L, attention shifts away from I to PL, because attending to I yields the wrong outcome. With PL attended, and I ignored, people then learn a strong association for PL to L. Notice that the learned attentional allocation depends on context. In the context of PE, attention is not shifted away from I, but in the context of PL, attention is shifted away from I.

Individual differences in selective attention

The magnitude of the highlighting effect varies across individuals. If the magnitude corresponds with the degree of attentional shifting, and if the degree of attentional shifting is a fairly stable individual characteristic, then the magnitude of highlighting ought to correlate with the magnitude of other phenomena attributed to selective attention in learning.

This prediction has been confirmed for two other measures of attention (Kruschke, Kappenman, & Hetrick, 2005). One measure is the magnitude of "blocking", which is another sort of response preference in associative learning. The left side of Table 1 shows that in blocking, late training consists of equal number of cases of A.B→X and C.D→Y. The only difference between them is training in the previous, early phase. A.B→X is preceded by cases of A→X, wherein A by itself predicts X. This previous training with A alone apparently blocks learning about B, as assayed by subsequent tests with the conflicting cues B.D, for which people strongly prefer outcome Y. This weakened association from B can be explained, at least in part, by learned *in*attention

to B: When learning A.B→X, the person already knows that A indicates X, so it is helpful to learn to suppress the distracting cue B. Now that blocking has been described, here's the point: Across individuals, the magnitude of blocking is correlated with the magnitude of highlighting. Moreover, eye tracking reveals that the magnitude of differential gaze at the cues, during test, is correlated with the magnitudes of highlighting and blocking (Kruschke et al., 2005; Wills, Lavric, Croft, & Hodgson, 2007).

Learned attention perseverates into subsequent learning

If people learn to attend to some cues or dimensions while suppressing attention to other cues or dimensions, then it is natural to suppose that the learned attention should perseverate into subsequent training even if the dimension values and/or the category assignments change. In particular, if the same dimension remains relevant after the change, then relearning should be easier than if a different dimension becomes relevant. This prediction, that an intra-dimensional shift should be easier than an extra-dimensional shift, has been well established in classical discrimination learning, especially in situations in which the cue values change when the relevance shifts (for a review, see, e.g., Slamecka, 1968).

Figure 1 shows a type of shift design that is quite different from traditional designs and solves some of their problems. Instead of changing the stimuli or outcomes when the relevance changes, all the stimuli and outcomes stay the same; only the mapping between them changes. Therefore there is no novelty in the stimuli or outcomes to indicate a change across which knowledge must be transferred. Specifically, learners first were trained on the Type II structure shown in the upper middle of Figure 1. Then they were seamlessly shifted to one of the structures shown in the right column of Figure 1. Both of the right structures have only a single relevant dimension. In the upper right (labeled "Relevant Shift"), the newly-relevant dimension is one of the dimensions that was previously relevant for the Type II structure. In the lower right (labeled "Irrelevant Shift"), the newly relevant dimension is the dimension that was irrelevant in the previous Type II structure. Notice that in both shifts there are exactly four stimuli that change their outcomes, and therefore any difference in difficulty of shift cannot be attributed to how many stimulus-outcome correspondences must be relearned. Finally, notice that the Type-II structure of the initial phase makes all the dimensions have zero correlation with the outcome. In other words, for any single dimension, there is 50% probability of both outcomes at both values of the dimension. Therefore, any difference in difficulty of shift cannot be attributed to the correlation between the dimension and the outcome in the initial phase. This type of structure has been used in subsequent studies by George and Pearce (1999) and by Oswald et al. (2001).

Results from human learners showed that the relevant shift was much easier to learn than the irrelevant shift (Kruschke, 1996b). This result is naturally explained by positing learned attention: People learned to attend to the two relevant dimensions for the Type II structure, and to ignore its irrelevant dimension. Therefore, in subsequent training, it was relatively difficult to learn to attend to the previously irrelevant dimension.

Analogous results have been obtained for learning after highlighting and after blocking. When trained on new associations involving previously highlighted cues, learning is faster (Kruschke, 2005). When trained on new associations involving previously blocked cues, learning is slower (Kruschke & Blair, 2000; Kruschke, 2005; Le Pelley & McLaren, 2003; Le Pelley, Oakeshott, Wills, & McLaren, 2005). Again, these results are naturally explained in terms of learned selective attention.

Competition for attention explains effects of cue salience

If attention has a role in associative learning, then cue salience should have an effect in learning, because salience connotes attraction of attention. The term "salience" has no generally accepted definition, but salience is defined here as the relatively long-lived ability of a cue to attract attention. A cue's salience might be assessed by its initial attentional allocation at the beginning of an experiment, before additional learning has shifted attention. The salience of a cue is always relative to the saliences of other cues that are present at the same time.

As one example, consider a situation in which words presented on a computer screen must be associated with corresponding key presses. The word "peek" might indicate pressing the F key, while the word "toll" indicates pressing the J key. The correspondence is probabilistic, however. Suppose that in addition to those words that are correlated with the correct key press, the screen also displays other words that are not correlated with the correct key press. The learner does not know in advance which words are correlated or uncorrelated with the correct answer. If the words compete for attention during learning, then learning about the relevant cues should be influenced by the relative salience of the irrelevant words. If the irrelevant words are highly salient, such as "boy" and "cat", then learning about the relevant words should be relatively difficult. If the irrelevant words are obscure and not salient, such as "nabob" and "witan", then learning about the relevant words should be relatively easy. This prediction, which is novel from a specific attentional theory of associative learning, was confirmed in Experiment 4 of Kruschke and Johansen (1999). For this example, involving words, salience refers to the encodability (e.g., concreteness) and associability (e.g., meaningfulness). In other applications, salience has been instantiated as the intensity or quantity of the cue. For example, when the cue is the color red, then the cue salience might be manipulated by the amount or density of red pixels on the computer screen (Denton & Kruschke, 2006).

As another example, recall the training procedure for blocking (left side of Table 1). The redundant cue B, added to the already-learned cue A, was learned to be ignored because it (cue B) distracted attention from the diagnostic cue. Thus,

blocking is caused, at least in part, by learned inattention to the blocked cue. This theory predicts that the degree of blocking should be modulated by the relative salience of the to-be-blocked cue. In particular, if cue B is highly salient, then it should be difficult to block. Indeed, when cue B is highly salient, it might even dominate learning in the second phase, actually robbing cue A of some control over responding. This prediction has been confirmed in humans and animals (Denton & Kruschke, 2006; Hall, Mackintosh, Goodall, & Dal Martello, 1977).

Attention can shift between representations

The previous sections have assumed that attention can be allocated to present/absent cues, such as the word "cat", or to values of dimensions, such the specific colors blue and red, or to entire dimensions, such as height. But attention can also be allocated to representational systems, such as a rule system or an exemplar system. The idea is that learners can associate stimuli with outcomes via a variety of different types of mappings. Some mappings might be mediated by exemplars. In an exemplar system, if a stimulus is similar to exemplars in memory, then the system anticipates the outcome stored in those exemplars. Other mappings might be mediated by rules. In a rule system, if a stimulus satisfies a specific condition, then the system anticipates the corresponding outcome. The condition for a rule typically spans a much larger area of stimulus space than an exemplar. For example, a rule might have as its condition, "anything taller than 3 cm", whereas an exemplar might have as its condition, "something very nearly 3 cm tall and 2 cm wide and colored green and weighing more than a kilogram". As another example, in learning to map continuous cues to continuous outcomes (i.e., in so-called *function learning*), an exemplar system would map a cue value near $x = 3.0$ (say) to an outcome of $y = 6.0$ (say), but a rule system could map *any* value of x to $y = 2x$. Learners should allocate attention to the different types of mappings according to how efficiently and accurately the mappings accommodate the training items. The allocation of attention among representational systems is learned, and the tuning of the mappings within the systems is learned.

Empirical evidence for this sort of attentional allocation has come from a series of experiments in category learning. The stimuli for these experiments had two continuous dimensions, like the Filtration and Condensation stimuli on the left of Figure 1. The structure of the categories was much like the Filtration structure, for which the stimuli were accurately classified by a simple rule: If the stimulus has Dimension A value left of center, then the stimulus is in category 1. But the structure was more involved than the simple Filtration structure, because it also had exceptions to the rule. The exceptions were arranged so that extrapolation beyond the training cases could be tested. The attentional theory predicts that for test cases that are fairly far from the exceptions, but even farther from the trained rule cases, responses should nevertheless favor the rule-predicted outcomes. This prediction stems from how attention is allocated to the systems. Attention goes to the exemplar system especially when the stimulus is highly similar to a known exception, but otherwise attention may prefer the rule system, which accommodates most of the training items. This prediction and many others have been confirmed in a series of experiments (Denton, Kruschke, & Erickson, 2008; Erickson & Kruschke, 1998, 2002; Kruschke & Erickson, 1994). Additional evidence comes from the theory predicting switch costs, i.e., response time increases, when stimuli switch across trials from being rule-mapped to being exemplar-mapped, and vice-versa (Erickson, 2008). Other category learning experiments, using various other structures and stimuli, along with extensive explorations of models, have bolstered interpretations in terms of distinct representational subsystems that are allocated attention in different circumstances (Lewandowsky, Roberts, & Yang, 2006; Little & Lewandowsky, 2009; Yang & Lewandowsky, 2003, 2004).

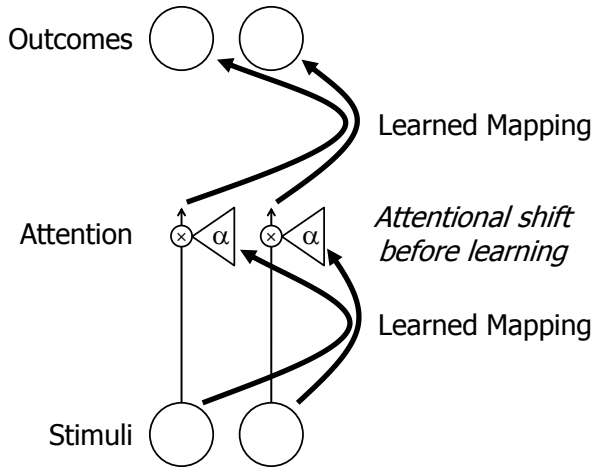

Figure 2. General framework for models of attentional learning. The stimuli are represented at the bottom of the diagram by activations of corresponding nodes. Thick curved arrows denote learned associative mappings. Attention is denoted by "α" in the middle layer and acts as a multiplier on the stimuli. On a given trial of learning, attention is shifted before the mappings are learned.

General Framework for Models of Attentional Learning

The phenomena reviewed in the previous section share an explanatory framework in which attention is rapidly reallocated across cues, dimensions, or representations. The learning of associations depends on the re-allocation of attention, and the re-allocation is itself learned. Figure 2 diagrams this general framework as an associative architecture. Associations feed forward from the stimuli at the bottom of the diagram to the outcomes at the top of the diagram.

In the general framework of Figure 2, attention is thought of as multiplicative gates on the stimulus components. The attentional gate on a stimulus component is indicated in Figure 2 by the multiplication sign and the α symbol accom-

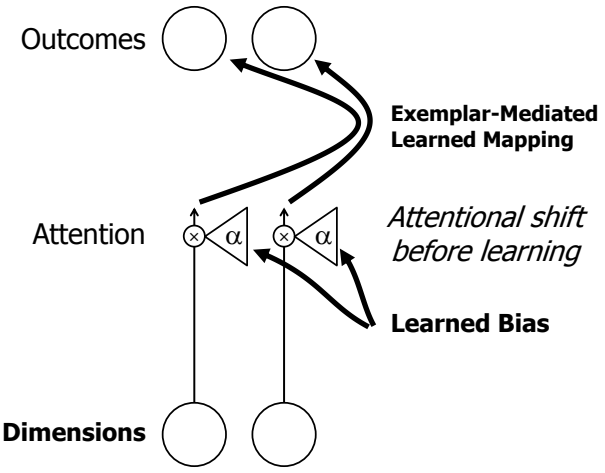

Figure 3. The RASHNL model, successor to ALCOVE. The stimuli are assumed to be values on dimensions, and attention is allocated to dimensions. The learned attentional allocation is simplistically assumed to be a bias that applies equally to all stimuli, unlike the general framework which allows stimulus-specific attentional mappings. The mapping from attended stimuli to outcomes is mediated by exemplars.

panying each stimulus component. The attentional values are shifted in response to feedback regarding correct outcomes. Attention is shifted away from stimulus components that generate error, toward stimulus components that reduce error. The re-allocated attention values are then learned, as associations from the stimuli. The outcomes are learned as associations from the attentionally filtered stimuli.

The general framework can be instantiated with different specific formalisms, depending on the situation to be modeled and the complexity demanded by the data. For example, the mappings between layers might be accomplished by simple linear associators if the domain and behavior are simple enough. But a more general model would need more complex representational options to accommodate more complex, non-linear mappings. The remainder of the chapter reviews several specific formal instantiations of the general framework in Figure 2.

Particular Instantiations

Learned attention across exemplars: RASHNL / ALCOVE

Figure 3 shows the RASHNL model (Kruschke & Johansen, 1999), successor to the ALCOVE (Kruschke, 1992) model. ALCOVE was a connectionist implementation of the Generalized Context Model (GCM; Nosofsky, 1986), which in turn was a generalization of the Context Model (Medin & Schaffer, 1978). RASHNL is an acronym for "Rapid Attention SHifts aNd Learning". The name is a play on the word "rational" because the model mechanisms are all driven by the rational goal of error reduction, even though the behavioral results are rash attentional shifts and seemingly irrational generalization behaviors. The name also gives a nod to the Rational Model of categorization by Anderson (1991).

In the RASHNL model (and ALCOVE), stimuli are represented by values on psychological dimensions, as denoted at the bottom of Figure 3. For example, the stimulus might have a height of 27 and a brightness of 12. The stimulus coordinates, e.g., $\langle 27, 12 \rangle$, are compared to the coordinates of exemplars in memory. The memory exemplars are activated to the extent that they are similar to the presented stimulus. Thus, memory exemplars that are very similar to the stimulus are strongly activated, but memory exemplars that are highly dissimilar to the stimulus are only weakly activated. The exemplars then propagate their activation to the outcome nodes along weighted connections. The weights encode the learned degree of association between each exemplar and the outcomes. This exemplar mediation of the mapping to outcomes from stimulus dimensions is indicated in the upper right of Figure 3.

The key role for attention is how much each stimulus dimension is used in the calculation of similarity. In the Filtration structure of Figure 1, for example, dimension A will be strongly emphasized when determining the similarity of the stimulus to the exemplars, while dimension B will be mostly disregarded as irrelevant.

When a stimulus is first presented, attention is allocated to its dimensions according to previously learned biases, as shown in the lower right of Figure 3. Activation is propagated up to the outcome nodes, to determine a predicted outcome. The prediction is a magnitude of preference for each outcome. Then corrective feedback is supplied, and the discrepancy between the predicted outcome and the actual outcome is computed. This prediction error is used to drive all learning in the model.

The first response to the error is a rapid reallocation of attention. This rapid shift is denoted in the middle right of Figure 3. Attention is shifted toward dimensions that reduce error, and away from dimensions that cause error. For example, when learning the Filtration structure of Figure 1, error is reduced by decreasing attention on Dimension B, because the collapsing of Dimension B brings closer together the exemplars that map to the same outcome, whereby learning about one exemplar enhances generalization to other exemplars.

After the attention to dimensions has been reallocated, then the model attempts to retain the reallocation for future use. This learning of the reallocation is stored in the bias weights. This learning is error driven, just like the initial rapid reallocation. The bias weights are adjusted to reduce the discrepancy between the initial attentional allocation and the reallocation demanded by the actual outcome. It is important to understand that the rapid shift of attention, in response to corrective feedback, is distinct from the learning of that shift. The shift might be large, but the large shift might not be retained to the next trial if the shift is not learned. Figure 3, like the general framework in Figure 2, points out this distinction by the label, "attentional shift before learning". The ALCOVE model, which was the precursor to RASHNL, does not have a rapid reallocation of attention before the learning of attention.

At the same time that the attentional shift is learned, the model attempts to learn the correct outcomes, by adjusting associations between the activated exemplars and the outcome nodes. An associative weight between an outcome and an exemplar is adjusted only to the extent that the exemplar is activated and there is error at the outcome node.

This sort of model has been shown to accurately fit learning performance and generalization behavior in a number of situations. For example, when applied to the Filtration and Condensation structures in Figure 1, ALCOVE shows a robust advantage for Filtration (Kruschke, 1993). When applied to the Type II and Type IV structures in Figure 1, ALCOVE again shows accurate fits to human learning data (Kruschke, 1992; Nosofsky et al., 1994). When applied to the Relevant and Irrelevant shifts in Figure 1, ALCOVE exhibits a strong advantage for the Relevant shift (Kruschke, 1996b). When applied to situations in which irrelevant cues have different saliences, as with the words "cat", "toll", and "witan" described earlier, the RASHNL model nicely captures human utilizations of the cues (Kruschke & Johansen, 1999). The RASHNL model, with its rapid shifts of attention, also qualitatively mimics the large individual differences in attentional allocation seen in human learners (Kruschke & Johansen, 1999). RASHNL has also used been used in clinical assessment. For example, men's attention to women's facial affect or body exposure, and other socially-relevant cues, were studied by Treat and colleagues (Treat, Kruschke, Viken, & McFall, 2010; Treat, McFall, Viken, & Kruschke, 2001; Treat et al., 2007). Male participants learned arbitrary category labels assigned to photos of women. Fits of RASHNL to the learning data revealed that these two stimulus dimensions were difficult to selectively attend, and, in particular, that learning to re-allocate attention away from the initial individual biases was very difficult.

Although RASHNL is a successor to ALCOVE, RASHNL has not yet been tested on all the data sets for which ALCOVE has been tested. ALCOVE is nearly (but not exactly) a special case of RASHNL for which the magnitude of rapid attention shifting is set to zero, so in that sense RASHNL might trivially be able to reproduce the behaviors of ALCOVE. More challenging, however, would be to simultaneously fit the data that motivated the creation of RASHNL and the data that have been successfully fit by ALCOVE. For example, it is not known whether RASHNL could reproduce the advantage of Relevant over Irrelevant shifts that ALCOVE shows robustly. Presumably the answer is yes, because attentional learning can be slow even when initial shifts are large, but simulations have yet to be conducted.

Context-specific learned attention: EXIT / ADIT

Figure 4 shows the EXIT model (Denton & Kruschke, 2006; Kruschke, 2001a, 2001b; Kruschke et al., 2005), successor to ADIT (Kruschke, 1996a). Ideas similar to those formalized by ADIT were previously described informally by Medin and Edelson (1988). The name ADIT is an acronym for Attention to Distinctive InpuT. The word "adit" refers to an entrance to a place that is mined, and so the play on words is that the model is an entrance to the mind. The successor model, EXIT, was so named because an exit can come after an adit.

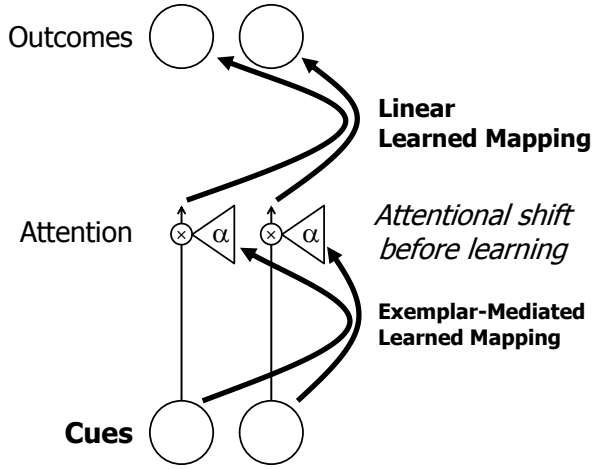

Figure 4. The EXIT model, successor to ADIT. The stimuli are assumed to be present/absent cues. The mapping from cues to attention is exemplar-mediated, whereas the mapping from attentionally gated cues to outcomes is simplistically assumed to be linear.

EXIT is another instantiation of the general framework of Figure 2. The EXIT model assumes that stimuli are represented as present/absent cues (instead of values on dimensions, as in RASHNL). Most importantly, EXIT assumes that the learned attentional allocation can be exemplar specific. Thus, attention might be allocated one way for some stimuli, but a different way for other stimuli. RASHNL and ALCOVE do not have this flexibility. The allocation of attention in EXIT is a learned mapping from stimulus cues to attentional values, mediated by exemplars. A comparison of EXIT in Figure 4 and RASHNL in Figure 3 shows this contrast in the associative weights going into the attention gates.

Because EXIT has been applied to relatively simple mappings of stimuli to outcomes, the model incorporates only a linear associator to outcomes from attended cues. This use of a linear associator on the outcome layer is merely for simplicity and reduction of parameters. If the model were applied to more complex situations, a more complex associator would have to be used, such as the exemplar-mediated associator used in RASHNL. This option is discussed more in a later section.

Aside from the representational differences just described, processing in EXIT proceeds much like processing in RASHNL. When a stimulus appears, activation is propagated up the network and an outcome is predicted. Corrective feedback is provided, and the degree of error is computed. At this point, attention is rapidly reallocated across the cues, away from cues that cause error, to cues that reduce error. This re-allocated attention serves as a target for error-driven subsequent learning of associations from stimulus cues to attentional allocation. As in RASHNL, attention shifting and attention learning are two distinct steps of processing. In

particular, it is possible to have shifting without learning of the shifts, as was assumed in the ADIT model.

EXIT (and ADIT) have been shown to fit many complex data sets from human learning experiments, including the highlighting and blocking designs in Table 1 (Denton & Kruschke, 2006; Kruschke, 1996a, 2001a, 2001b, 2005; Kruschke et al., 2005). EXIT is especially good at capturing detailed choice preferences for a variety of cue combinations tested after highlighting. EXIT accomplishes this mimicry of human responding by the way the model shifts and learns attentional allocation to cues, especially during learning of the late-phase cases of I.PL→L (see Table 1). When a case of I.PL→L appears, attention is allocated away from cue I, because it is already associated with the other outcome E, and attention is shifted toward cue PL, because it does not conflict with the correct outcome. The attentional reallocation is then learned, specific to the cue combination I.PL. In other words, EXIT learns that when stimuli I.PL are presented, suppress attention to I and attend to PL, but when stimuli I.PE are presented, maintain some attention to both cues.

Because EXIT has learned an attentional allocation for particular cues, this learning will persist into subsequent phases of training. In particular, if subsequent phases of training have stimulus-outcome mappings with the same relevant cues, then subsequent learning should be easy. But if subsequent training has a stimulus-outcome mapping with different relevant cues, then the subsequent learning should be relatively difficult. These predictions have been confirmed and modeled by EXIT (Kruschke, 2005). For example, after highlighting, participants continued into subsequent training for which two novel outcomes were perfectly predicted by two cues that played that roles of the imperfect predictors in the previous highlighting phases. For some participants, the I cues were accompanied by the PE cues from the previous highlighting phases. When accompanied by PE, the I cues should receive some attention, and therefore the new learning should be relatively easy. Other participants learned about the I cues accompanied by PL cues from the previous highlighting phases. When accompanied by PL, the I cues should receive less attention, and therefore the new learning should be relatively difficult. This prediction was confirmed, and modeled by EXIT. Analogous results have been shown and modeled for blocking (Kruschke & Blair, 2000; Kruschke, 2001b, 2005).

EXIT and ADIT have also been shown to accurately fit human choice preferences for a variety of probabilistic mappings, and structures involving differential base rates of the categories. One of the benefits of the attentional interpretation provided by the models is that some of the empirical phenomena can be better understood. In particular, some findings regarding differential base rates could be reinterpreted in terms of attention shifts caused by the induced order of learning (Kruschke, 1996a, 2010). When two categories have very different base rates, the high-frequency category is learned first. The low-frequency category is subsequently learned, and attention is re-allocated during learning of the low-frequency category. The attentional re-allocation accounts for many choice preferences that are otherwise theoretically perplexing.

In general, attentional re-allocation is beneficial for fast learning. For example, when learning I.PL→L in the late phase of highlighting, the shift of attention away from cue I protects the previously learned association from I to E, thereby retaining that association for accurate future prediction of outcome E. Learning of the shift also prevents subsequent errors on when I.PL are presented again, because attention is allocated away from cue I. Thus, attentional allocation is a mechanism by which an organism can accelerate learning of new items while retaining knowledge of old items. This argument has been suggested in several previous publications (e.g., Kruschke & Johansen, 1999; Kruschke, 2003c). Recently it has been shown that when connectionist architectures are evolved using simulated genetic algorithms, the optimal architectures are ones, like EXIT (Kruschke & Hullinger, 2010), that have the ability to rapidly reallocate attention across cues. In the simulated evolution, the only adaptive pressure put on the evolving learners was to learn fast, i.e., to have as little total error as possible during the lifetime of the organism. One of the key structural aspects of the training environment was that some cues changed less frequently than others. The slowly changing cues formed the context for the rapidly changing cues, and the attentional reallocation took advantage of changing cues relevances in different contexts.

Attentionally modulated exemplars and exemplar-mediated attention

A comparison of RASHNL (Figure 3) and EXIT (Figure 4) invites two natural generalizations. First, the present/absent cues assumed by EXIT might be generalized to dimensions, as assumed in RASHNL. Such a generalization has been explored, wherein a dimensional value is represented by thermometer-style encoding of present/absent elements that represent levels on the dimension (Kalish, 2001; Kalish & Kruschke, 2000). In these models, attention can be allocated to different dimensions, and also to different values within dimensions. It remains to be seen whether or not this approach will be useful as a general solution to representing stimuli and selective attention to their components.

Second, a general model would allow exemplar-mediated mapping at both the attentional and outcome levels. Such a generalization has been developed and reported at conferences (Kruschke, 2003a, 2003b), but not yet published. In this generalized model, the exemplars that mediate the outcome mapping are attentionally modulated, as in RASHNL. In addition, there are exemplars that mediate the attention mapping, as in EXIT. Hence the model has attentionally modulated exemplars and exemplar-mediated attention. The attentional modulation on the exemplars also determines whether or not new exemplars are recruited during learning: If attention shifts away from candidate exemplars, they are not retained.

A key aspect of this generalization is that exemplars for the outcome-mapping encode both the cues and the attention paid to them. This is different than the type of exemplars

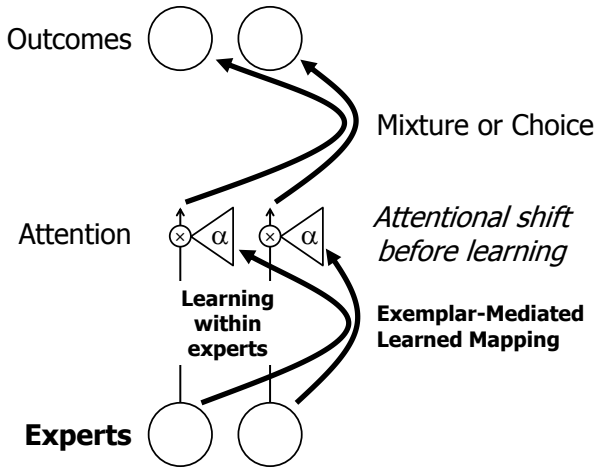

Figure 5. The ATRIUM and POLE models, both instances of a mixture-of-experts architecture. Attention is allocated to "experts", each of which constitutes a complete learning system from stimuli to outcomes. The mapping from stimuli to attention is exemplar-mediated, so that different experts can dominate in different contexts. The internal details of the experts are not shown. The final outcome is a mixture or choice from among the attended experts.

used in RASHNL and ALCOVE. Instead, these generalized exemplars record the cues and their attentional gating. In other words, an exemplar for the outcome mapping contains not only the stimulus coordinates, but it also contains the attentional allocation when that stimulus is processed. An interesting consequence of this representation is that the attentional values that are stored in the exemplars can be adjusted to reduce error, and this error-driven adjustment of exemplar-specific attention accounts for so-called "retrospective revaluation" effects. Retrospective revaluation occurs when new learning occurs for a cue even when it is not present in the stimulus. In the generalized model, an exemplar node that encodes a present cue can retrospectively reduce or increase its stored attention to that cue, even when that cue is not present in the current stimulus. This approach to retrospective revaluation is promising, but was eclipsed by Bayesian approaches, discussed in a subsequent section. Nevertheless, the generalized exemplar approach deserves renewed attention in the future.

Mixture of Experts: ATRIUM / POLE

As mentioned in the introduction, attention can be allocated to representational systems, in addition to cues or dimensions or values on dimensions. For example, one representational system may map stimuli to outcomes via exemplar memory, while another representational system may map stimuli to outcomes via condition-consequent rules.

Figure 5 shows the basic structure of models that learn to allocate attention among "expert" subsystems. Each expert learns its own mapping from stimuli to outcomes, using its own form of representation. Which expert takes responsibility, for learning and responding to any particular stimulus, is determined by a gating network. The gating network is indicated in Figure 5 as an exemplar-mediated, learned mapping. As is the case for all the models described so far, the allocation of responsibility to experts is driven by error reduction: Attention is allocated to experts that accommodate the current training case, and attention is allocated away from experts that cause error. These models are cases of a mixture-of-experts architecture (Jacobs, Jordan, Nowlan, & Hinton, 1991; Jacobs, Jordan, & Barto, 1991).

The ATRIUM model has been used to model learning of categories in which the structure can be intuitively described as a rule with exceptions (Erickson & Kruschke, 1998; Kruschke & Erickson, 1994). As mentioned in the introduction, one of the key findings is that when people are tested with novel cases that demand extrapolation away from the training cases, responses are usually consistent with the rule, despite the fact that the nearest training case was an exception to the rule. Although some researchers have challenged the generality of the findings or shown that a variation of an exemplar-only model can capture some aspects of the findings (Nosofsky & Johansen, 2000; Rodrigues & Murre, 2007), follow-up experiments have repeatedly demonstrated that the effect is robust and that exemplar-only models cannot accommodate extrapolation of rules with exceptions, while the mixture-of-experts ATRIUM model can (Denton et al., 2008; Erickson & Kruschke, 2002). Thus, rule and exception learning continues to challenge exemplar-only models, and mixture-of-expert models continue to better fit the data. Results from other category structures also point to the utility of mixture-of-expert models (Lewandowsky et al., 2006; Little & Lewandowsky, 2009; Yang & Lewandowsky, 2003, 2004).

Analogous findings have been found in function learning. In function learning, instead of mapping stimuli to categorical outcomes, both the stimuli and outcomes are metric. Typically some simple function relates the input to output values, such as a linear or low-order polynomial. Consider a situation in which most of the training cases follow a simple linear function, but a few exceptions deviate from the line. In particular, one of the exceptions is the most extreme of the trained stimulus values. When tested for extrapolation beyond this last training case, people tend to revert to the rule, rather than respond according to the nearest (exception) exemplar. Several cases of this sort of behavior were reported and modeled by Kalish, Lewandowsky, and Kruschke (2004). The model was a mixture-of-experts architecture called POLE (Population Of Linear Experts) in which expert modules for simple linear functions were gated along with exemplar-based experts. The model learned to allocate responsibility to the linear expert except when the stimulus was very similar to one of the learned exceptions.

Finally, the classic attentional model of Mackintosh (1975) can be generalized and re-expressed in a mixture-of-experts framework (Kruschke, 2001b). Each expert consists of a single distinct cue, trying to learn on its own to predict the presence or absence of the single outcome (i.e., the unconditioned stimulus in animal conditioning experiments). The attentional gate allocates responsibility to the cues to the

extent that they successfully predict the outcome.

Locally Bayesian Learning

The previous instantiations of the general framework have all relied on error-driven learning, i.e., gradient descent on error as in backpropagation (Rumelhart, Hinton, & Williams, 1986). In all those models, the knowledge of the model at any time consists of a single specific combination of associative weights. There is no representation of other weight combinations that might be nearly as good.

An alternative formalization of learning comes from a Bayesian approach. In a Bayesian learner, multiple possible hypotheses are entertained simultaneously, each with a learned degree of credibility. For example, consider a situation with two cues, labeled A and B, and a single outcome X. The Bayesian learner might entertain three possible hypotheses: $H_{A(\neg B)}$: A, but not B, indicates X; $H_{B(\neg A)}$: B, but not A, indicates X; and, $H_{A \vee B}$: A or B indicate X. If the Bayesian learner experiences training cases of A→X, then the credibilities of $H_{A(\neg B)}$ and $H_{A \vee B}$ increase, while the credibility of $H_{B(\neg A)}$ decreases. "Hypotheses" need not be expressed as logical rules. Indeed, most Bayesian models involve hypotheses about mappings that have continuous values, and the hypothesis space consists of the infinite space of all possible combinations of the continuous values. The fact that the hypothesis space is infinite does not mean that a Bayesian learner needs an infinite-capacity mind. On the contrary, the credibility of the infinite hypotheses can be summarized by just a few values, as, for example, the mean and standard deviation summarize an infinitely wide normal distribution. An infinite distribution can also be summarized approximately by a large representative sample.

Bayesian learning models have many attractions, both in terms of their general computational abilities and as models of mind. There is not space here to review their many applications, but an overview is provided by Chater, Tenenbaum, and Yuille (2006), and a tutorial of their application to associative models is provided by Kruschke (2008). One learning phenomenon that is easily explained by Bayesian models, but that is challenging for many non-Bayesian associative models, is backward blocking. In backward blocking, the training phases of the blocking procedure in Table 1 are run in backward order. Curiously, the blocking effect is still exhibited by human learners (e.g., Shanks, 1985; Dickinson & Burke, 1996; Kruschke & Blair, 2000). In Bayesian accounts of backward blocking, different combinations of associative weights are considered simultaneously, with more belief allocated to the combination that is most consistent with the training items. Because the cases of A→X decrease the credibility of $H_{B(\neg A)}$, as explained in the previous paragraph, cue B is effectively blocked regardless of when the cases of A→X occur (Dayan & Kakade, 2001; Tenenbaum & Griffiths, 2003).

A theoretical framework that combines the attentional and Bayesian approaches is called "locally Bayesian learning" (LBL, Kruschke, 2006b). The overall LBL framework is quite general and does not rely on any notion of attention.

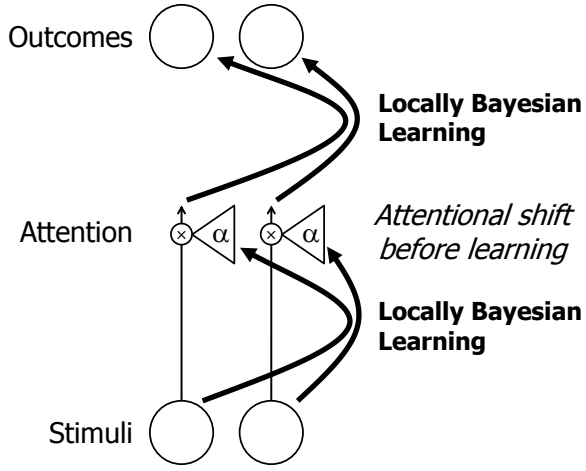

Figure 6. Locally Bayesian learning applies generally to componential models. Here it is applied to the case in which the first layer learns to allocate attention, and the second layer learns to generate an outcome, given the attended stimuli.

The general LBL framework is based on the idea that a learning system may consist of a sequence of subsystems in a feed-forward chain, each of which is a (locally) Bayesian learner. The argument for locally-learning layers was as follows. First, Bayesian learning is very attractive for explaining retrospective revaluation effects such as backward blocking, among many other phenomena (Chater et al., 2006). Second, Bayesian learning may also be *un*attractive for a number of reasons. In highly complex hypothesis spaces, it may be extremely difficult to accomplish even approximate Bayesian learning. In other words, keeping track of all viable hypotheses, even approximately, may be computationally intractable. Furthermore, many globally Bayesian models do not explain learning phenomena such as highlighting (Table 1) that depend on training order, because the models treat all training items as equally representative of the world to be learned, regardless of when the items occurred. Finally, the level of analysis for theories of learning is arbitrary: Learning occurs simultaneously at the levels of neurons, brain regions, functional components, individuals, committees, institutions, and societies, all of which may be modeled (in principle) as Bayesian learners. Therefore, a system of locally Bayesian learning components may retain some attractions of Bayesian models while also implementing Bayesian learning in smaller, tractable hypothesis spaces.

The general framework for locally Bayesian learning has been instantiated in a particular two-layer model, wherein one layer learns how to allocate attention to cues, and a second layer learns how to associate attended cues with outcomes (Kruschke, 2006b, 2006a). Figure 6 shows the structure of the model. When a stimulus is presented, the model generates a predicted outcome as follows. The mapping, from stimuli to attention allocation, comprises many candidate hypotheses regarding how to allocate attention. Each hypothesis has a degree of credibility at that point in train-

ing. The actual allocation of attention is taken to be the average allocation across all the hypotheses, weighed by their credibilities. The attentionally gated stimulus is then delivered to the next layer, which maps the stimulus to outcomes. This layer again consists of many candidate hypotheses, each with a degree of credibility. The predicted outcome is the average of the predicted outcomes across all the hypotheses, weighted by their credibilities.

When corrective feedback is presented, learning occurs. (There are different possible learning dynamics. The one described here is the one that most closely mimics the processing of previously described attentional models.) First, the top-most layer determines which allocation of attention would best match actual correct outcome. In other words, the system finds the allocation of attention that would be most consistent with the currently credible hypotheses of the outcome layer. This is the "rapid shift of attention before learning". This best allocation of attention is treated as the target for the lower layer. With the attention reallocated, both layers than learn according to standard Bayesian inference. In the attentional layer, credibilities are shifted toward hypotheses that are consistent with the current stimulus being mapped to the target allocation of attention. In the outcome layer, credibilities are shifted toward hypotheses that are consistent with the attended components being mapped to the correct outcomes.

This locally Bayesian learning generates behavior that depends on training trial order. The reason, that learning depends on trial order, is that the internal attentional target depends on the current beliefs in the outcome layer. In other words, a particular stimulus and actual outcome will be given different internal attentional targets, depending on the current credibilities of outcome hypotheses.

The highlighting effect (recall Table 1) is a trial-order effect. Highlighting occurs robustly even when the total frequency of each category is the the same (Kruschke, 2010). Highlighting occurs in the LBL model because of how attention is allocated in the late-training phase. By the time the late-training phase has occurred, the model's outcome layer has established some credibility in associations from cue I to outcome E. When the training case I.PL→L occurs, the model reallocates attention away from cue I, because leaving attention on I violates its current beliefs. Then the model learns, in its lower layer, that when I.PL occur, allocate attention away from I toward PL. The model's upper layer retains its belief that cue I is associated with outcome E.

Whereas it is difficult for many globally Bayesian models to account for highlighting, LBL can. In addition, unlike the non-Bayesian models such as EXIT, LBL can also account for backward blocking, as mentioned earlier. Backward blocking is accommodated by the upper layer in LBL, but backward blocking of cues to outcomes does not require Bayesian learning in the lower layer. Recent empirical work has provided suggestive evidence that there can also be backward blocking of cues that are themselves uncorrelated with the outcomes, but which indicate what other cues are relevant to the outcome. This backward blocking of cues to relevance can be accommodated by Bayesian learning in the lower layer of the LBL architecture (Kruschke & Denton, 2010).

The description of LBL has been informal, because the exact mathematical expression can take different forms. For example, Kruschke (2006b) implemented LBL with a small finite set of weight combinations in a nonlinear mapping. On the other hand, Kruschke and Denton (2010) implemented LBL with layers of Kalman filters, which represent infinite spaces of possible associative weight combinations in linear mappings. (An introduction to Kalman filters is provided in Kruschke, 2008) Both approaches show the same qualitative behavior.

In summary, LBL in general alleviates some of the computational problems of globally Bayesian learning, yet retains the attraction that learning may be Bayesian at some levels of analysis. LBL as specifically applied to attentional learning exhibits many human learning phenomena that are challenging to globally Bayesian models or to non-Bayesian models.

Relations to other models, and possible generalizations

The general attentional framework described in the previous sections has emphasized how attention shifts toward representational components (such as cues, dimensions, or expert modules) that accommodate the goal (such as accurately predicting the outcome), and how attention shifts away from representational components that conflict with, or are irrelevant to, the goal. In other words, the attentional mechanism pays attention to information that is useful, and ignores information that is useless. Attentional mechanisms effectively compress out redundancies in the stimulus encoding, simplifying the encoding of information, and minimizing the description length of conditions for generating responses. Attentional shifting might be construed as a mechanism for on-the-fly discovery of minimal encoding. Thus, the attentional approach has conceptual affinity with the simplicity model of Pothos and Chater (2002), as summarized by Pothos, Chater & Hines (Chapter 9 of this volume). Further suggestive of a close relation between attention and encoding simplicity is the fact that ALCOVE, which dynamically re-allocates attention, predicts the relative difficulties of different category structures much like (but not always exactly the same as) the ordering predicted by the theory of structural complexity and categorical invariance by Vigo (2006, 2009). Attention might also be construed as a mechanism that mimics or implements prior beliefs about candidate rule structures, as formalized in some Bayesian approaches to category learning (Goodman, Tenenbaum, Feldman, & Griffiths, 2008).

All of the instantiations of the general attentional framework in Figure 2 have intellectual ancestors and cousins. The ALCOVE model, in particular, is a direct connectionist implementation of the Generalized Context Model (GCM; Nosofsky, 1986), which is summarized by Nosofsky (Chapter 2 of this volume). ALCOVE extends the GCM by providing a mechanism whereby the dimensional attention strengths are learned (rather than freely estimated to fit data at different points in training) and the exemplar-to-

outcome associations are learned (rather than set equal to co-occurrence frequencies). One advantage of ALCOVE over the GCM is that ALCOVE provides a simple and intellectually appealing learning mechanism, with only two learning rate parameters (on attention and association weights) instead of several dimensional attention parameters. Some disadvantages of ALCOVE are that the learning mechanism might not mimic actual learning performance, and that predictions of ALCOVE can only be ascertained via trial-by-trial simulations of entire training sequences whereas predictions of the GCM do not require trial-by-trial training.

Exemplars in ALCOVE and RASHNL were, for convenience, assumed to be either (1) preloaded as a random covering of the stimulus space, or (2) preloaded because of prior exposure to, or inference about, the range of possible stimuli, or (3) recruited on-the-fly when novel stimuli occurred. A different mechanism for recruiting exemplars was used in the model described in the section of this chapter headed "attentionally modulated exemplars and exemplar-mediated attention". In that section it was mentioned that exemplars can mediate the mappings in both layers of the general framework, and the exemplar activations can themselves be attentionally modulated. When attention modulates the exemplars, the attentional modulation can govern whether candidate exemplars, elicited by stimuli on every trial, are retained for future learning or retired after that trial. Specifically, if attention is shifted toward a candidate exemplar (because previously learned exemplars cause error), it is retained, but if attention is shifted away from a candidate exemplar (because previously learned exemplars are already performing well), it is retired. Notice that this mechanism constitutes an error-driven attentionally-based exemplar recruitment mechanism. This recruitment mechanism may be similar in spirit to cluster recruitment in the supervised mode of the SUSTAIN model (Love, Medin, & Gureckis, 2004), summarized by McDonnell and Gureckis (Chapter 10 of this volume), but SUSTAIN uses an error threshold for recruitment instead of an attentional mechanism. SUSTAIN also recruits clusters during unsupervised learning, using a novelty threshold, which might be mimicked in the proposed model by attentional shifts in an auto-encoding architecture (in which the outcome layer includes a copy of the cues). The ability of SUSTAIN to create adjustable clusters is a representational advantage over exemplar-only models. Future models may benefit from combining the clustering ideas of SUSTAIN with other mechanisms of cluster recruitment.

Exemplar representation has also been the subject of extensive investigation by animal-learning researchers, albeit under the rubric of "configural" versus "elemental" representations. A prominent configural model in animal learning was created by Pearce (1994). The configural model is analogous to ALCOVE in many ways, such as having error-driven learning of associative weights between configural units and outcomes. But it differs from ALCOVE significantly by having no selective attention to cues, whereby different cues within a stimulus are selectively amplified or attenuated. Configural models are contrasted with elemental models, which attempt to account for learning by using stimulus representations only of individual cues, instead of combinations of cues. A prominent example of an elemental approach is the basic version of the Rescorla-Wagner model (without configural cues; Rescorla & Wagner, 1972), upon which the ADIT model is based. Other elemental models include the one by McLaren and Mackintosh (2000, 2002) which uses a representation akin to stimulus-sampling theory (SST; Atkinson & Estes, 1963; Estes, 1962) and is summarized by Livesey and McLaren (Chapter 7 of this volume). Another intriguing elemental model has been reported by J. A. Harris (2006), also based on SST but which includes a limited-capacity attention buffer. These elemental models, like Pearce's configural model, have no learned selective attention to cues. Future generalizations might profitably combine SST-like cue representations with learnable, selective attentional shifting, perhaps not unlike the manner explored by Kalish and Kruschke (2000).

The notion of learned selective attention to specific cues has a long heritage, however, in both animal and human learning research. As just one example from human-learning research, Medin and Edelson (1988) informally expressed ideas about attention shifting that were subsequently formalized in the ADIT and EXIT models. In the animal learning literature, the model of Mackintosh (1975) described how cue associabilities may adapt through training. Mackintosh's heuristic formalism was shown subsequently to be closely related to a mixture-of-experts model in which each expert is a single cue acting individually to predict the outcome (Kruschke, 2001b). The EXIT model instead sums the influences of the cues, thereby generating different behaviors (such as conditioned inhibition, see Kruschke, 2001b).

Both EXIT and ATRIUM use exemplars to mediate the mapping from input to attentional allocation. This allows the models to learn *stimulus-specific attentional allocation*. EXIT has stimulus-specific attention to cues, whereas ATRIUM has stimulus-specific attention to representational modules. Another prominent model that uses distinct representational modules is the COVIS model of Ashby, Alfonso-Reese, Turken, and Waldron (1998), summarized in its current form by Ashby, Paul and Maddox (Chapter 4 of this volume). ATRIUM and COVIS are analogous insofar as they both have rule-like and exemplar-like subsystems. Although both models are founded on the idea that people can learn category mappings via different representational subsystems, the formalizations of the idea have different motivations in the two models. ATRIUM was motivated by a unifying mathematical aesthetic, whereby the rule module and the exemplar module and the gating between them are all driven by the same mechanism: gradient descent on the overall error. COVIS was motivated by neuropsychological considerations, such that the rule (i.e., explicit verbal) subsystem has formalisms motivated by hypothesis testing, and the exemplar-like (i.e., procedural) system has formalisms motivated by neural mechanisms, and the competition between the subsystems is driven by a separate heuristic that combines the long-term accuracies of the modules (i.e., their "trust") with the decisiveness of each module regarding the current stimulus (i.e., their "confidence"). The module-combination

rule in COVIS is not stimulus-specific, but the module-gating mechanism in ATRIUM is stimulus-specific. This difference between the models generates different predictions for the category structure used by Erickson (2008).[1] That structure combined a one-dimensional rule, similar to the filtration structure in Figure 1, with an "information integration" structure, similar to the condensation structure in Figure 1, simultaneously in different regions of stimulus space. Consider stimuli that are near the boundary of the information-integration (i.e., condensation) training items, which are simultaneously far from any of the one-dimensional rule (i.e., filtration) training items. In ATRIUM, these stimuli will be given responses from information-integration categories, because the gating mechanism is stimulus-specific and therefore allocates attention to the exemplar module. In COVIS, on the other hand, these stimuli will be given responses corresponding to the one-dimensional rule, because the exemplar-like (procedural) module has very low confidence but the rule module has very high confidence. There are many other differences in details of the models, which future research may explore. Insights from the two models, and the phenomena to which they have been applied, might profitably be merged in a future generalization.

Essentially all the models mentioned in the preceding paragraphs learn by some form of error reduction. None of the models learns by applying Bayes' rule to the space of possible representations that could be learned. In a previous section of the chapter, the method of locally Bayesian learning (LBL) was described, wherein the learned attentional allocation and the learned mapping from (attentionally filtered) stimuli to outcomes are learned in a Bayesian fashion. The implementation of each layer in LBL can be any Bayesian associator. For example, Kruschke (2006b) used linear sigmoids with a finite space of weight combinations, but Kruschke and Denton (2010) used Kalman filters. Kruschke (2006b) suggested that the layers could instead be implemented by the (approximately Bayesian) Rational Model of Anderson (1991). Indeed, the fully Bayesian nonparametric approach of Sanborn, Griffiths, and Navarro (2006), summarized in its developed form by Griffiths, Sanborn, Canini, Navarro, and Tenenbaum (Chapter ** of this volume), could be used instead. Alternatively, the layers could learn about latent causes in a sigmoid-belief network, as proposed by Courville, Daw, and Touretzky (2006). All of these different representations incorporate the intellectual appeal of Bayesian rationality for local learning, along with the benefit of accounting for complex learning phenomena such as backward blocking, and the explanatory power of attentional learning.

Finally, the last few Bayesian models that were mentioned, such as nonparametric clustering and latent causes, are non-directional models: Unlike the feed-forward prediction of outcomes from stimuli, assumed by the diagrams throughout this chapter, these models simultaneously generate stimulus features and outcome features, and predict missing values from any other present values. (In the Bayesian literature, such models are referred to as "generative" as opposed to "discriminative".) Non-Bayesian connectionist models can also be designed to allow prediction of any missing features. Such models could involve feedback connections, as in the KRES model of Rehder and Murphy (2003) (cf. the model's recent generalization summarized by Harris and Rehder in Chapter 12 of this volume), or the models could instead use autoencoder-style architectures that have the complete stimulus pattern included in the output pattern to be predicted. This space of modeling possibilities is huge, but it is likely that accurate models of human learning will involve mechanisms for learned selective attention.

References

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*(3), 409–429.

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

Ashby, F. G., Paul, E. J., & Maddox, W. T. (2010). COVIS. In E. M. Pothos & A. J. Wills (Eds.), *Formal approaches to categorization* (chap. 4). Cambridge, UK: Cambridge University Press.

Atkinson, R. C., & Estes, W. K. (1963). Stimulus sampling theory. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology.* New York: Wiley.

Chater, N., Tenenbaum, J. B., & Yuille, A. (Eds.). (2006, July). Special issue: Probabilistic models of cognition. *Trends in Cognitive Sciences*, *10*(7), 287–344.

Courville, A. C., Daw, N. D., & Touretzky, D. S. (2006). Bayesian theories of conditioning in a changing world. *Trends in Cognitive Sciences*, *10*(7), 294–300.

Dayan, P., & Kakade, S. (2001). Explaining away in weight space. In T. Leen, T. Dietterich, & V. Tresp (Eds.), *Advances in neural information processing systems* (Vol. 13, pp. 451–457). Cambridge, MA: MIT Press.

Denton, S. E., & Kruschke, J. K. (2006). Attention and salience in associative blocking. *Learning & Behavior*, *34*(3), 285–304.

Denton, S. E., Kruschke, J. K., & Erickson, M. A. (2008). Rule-based extrapolation: A continuing challenge for exemplar models. *Psychonomic Bulletin & Review*, *15*(4), 780–786.

Dickinson, A., & Burke, J. (1996). Within-compound associations mediate the retrospective revaluation of causality judgements. *Quarterly Journal of Experimental Psychology: Comparative & Physiological Psychology*, *49B*, 60–80.

Erickson, M. A. (2008). Executive attention and task switching in category learning: Evidence for stimulus-dependent representation. *Memory & Cognition*, *36*(4), 749–761.

Erickson, M. A., & Kruschke, J. K. (1998). Rules and exemplars in category learning. *Journal of Experimental Psychology: General*, *127*(2), 107–140.

Erickson, M. A., & Kruschke, J. K. (2002). Rule-based extrapolation in perceptual categorization. *Psychonomic Bulletin & Review*, *9*(1), 160-168.

Estes, W. K. (1962). Learning theory. *Annual review of psychology*, *13*(1), 107–144.

George, D. N., & Pearce, J. M. (1999). Acquired distinctiveness is controlled by stimulus relevance not correlation with reward. *Journal of Experimental Psychology: Animal Behavior Processes*, *25*(3), 363–373.

[1] This implication was pointed out by Michael A. Erickson (personal communication, October 10, 2009).

Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive Science*, *32*(1), 108–154.

Griffiths, T. L., Sanborn, A. N., Canini, K. R., Navarro, D. J., & Tenenbaum, J. B. (2010). Nonparametric Bayesian models of categorization. In E. M. Pothos & A. J. Wills (Eds.), *Formal approaches to categorization* (chap. **). Cambridge, UK: Cambridge University Press.

Hall, G., Mackintosh, N. J., Goodall, G., & Dal Martello, M. (1977). Loss of control by a less valid or by a less salient stimulus compounded with a better predictor of reinforcement. *Learning and Motivation*, *8*, 145–158.

Harris, H. D., & Rehder, B. (2010). Knowledge and resonance in models of category learning and categorization. In E. M. Pothos & A. J. Wills (Eds.), *Formal approaches to categorization* (chap. 12). Cambridge, UK: Cambridge University Press.

Harris, J. A. (2006). Elemental representations of stimuli in associative learning. *Psychological review*, *113*(3), 584–605.

Jacobs, R. A., Jordan, M. I., & Barto, A. (1991). Task decomposition through competition in a modular connectionist architecture: the what and where vision tasks. *Cognitive Science*, *15*, 219–250.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., & Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural Computation*, *3*, 79–87.

Kalish, M. L. (2001). An inverse base rate effect with continuously valued stimuli. *Memory & Cognition*, *29*(4), 587–597.

Kalish, M. L., & Kruschke, J. K. (2000). The role of attention shifts in the categorization of continuous dimensioned stimuli. *Psychological Research*, *64*, 105–116.

Kalish, M. L., Lewandowsky, S., & Kruschke, J. K. (2004). Population of linear experts: Knowledge partitioning and function learning. *Psychological Review*, *111*(4), 1072–1099.

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

Kruschke, J. K. (1993). Human category learning: Implications for backpropagation models. *Connection Science*, *5*, 3–36.

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

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

Kruschke, J. K. (2001a). The inverse base rate effect is not explained by eliminative inference. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *27*, 1385–1400.

Kruschke, J. K. (2001b). Toward a unified model of attention in associative learning. *Journal of Mathematical Psychology*, *45*, 812-863.

Kruschke, J. K. (2003a). *Attentionally modulated exemplars and exemplar mediated attention.* Invited talk at the Seventh International Conference on Cognitive and Neural Systems, Boston University, May 28-31.

Kruschke, J. K. (2003b). *Attentionally modulated exemplars and exemplar mediated attention.* Keynote Address to the Associative Learning Conference, Gregynog (University of Cardiff) Wales, April 15-17.

Kruschke, J. K. (2003c). Attentional theory is a viable explanation of the inverse base rate effect: A reply to Winman, Wennerholm, and Juslin (2003). *Journal of Experimental Psychology: Learning, Memory and Cognition*, *29*, 1396-1400.

Kruschke, J. K. (2005). Learning involves attention. In G. Houghton (Ed.), *Connectionist models in cognitive psychology* (pp. 113–140). Hove, East Sussex, UK: Psychology Press.

Kruschke, J. K. (2006a). Locally Bayesian learning. In R. Sun (Ed.), *Proceedings of the 28th annual conference of the Cognitive Science Society* (pp. 453–458). Mahwah, NJ: Erlbaum.

Kruschke, J. K. (2006b). Locally Bayesian learning with applications to retrospective revaluation and highlighting. *Psychological Review*, *113*(4), 677–699.

Kruschke, J. K. (2008). Bayesian approaches to associative learning: From passive to active learning. *Learning & Behavior*, *36*(3), 210-226.

Kruschke, J. K. (2010). Attentional highlighting in learning: A canonical experiment. In B. Ross (Ed.), *The psychology of learning and motivation* (Vol. **, pp. **–**). **: Academic Press. (Pre-print available at author's website, http://www.indiana.edu/~kruschke)

Kruschke, J. K., & Blair, N. J. (2000). Blocking and backward blocking involve learned inattention. *Psychonomic Bulletin & Review*, *7*, 636-645.

Kruschke, J. K., & Denton, S. E. (2010). Backward blocking of relevance-indicating cues: Evidence for locally Bayesian learning. In M. E. LePelley & C. J. Mitchell (Eds.), *Attention and learning* (pp. **–**). **: **.

Kruschke, J. K., & Erickson, M. A. (1994). Learning of rules that have high-frequency exceptions: New empirical data and a hybrid connectionist model. In *The proceedings of the sixteenth annual conference of the cognitive science society* (pp. 514–519). Hillsdale, NJ: Erlbaum.

Kruschke, J. K., & Hullinger, R. A. (2010). The evolution of learned attention. In N. Schmajuk (Ed.), *Computational models of associative learning* (pp. **–**). **: **.

Kruschke, J. K., & Johansen, M. K. (1999). A model of probabilistic category learning. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *25*(5), 1083–1119.

Kruschke, J. K., Kappenman, E. S., & Hetrick, W. P. (2005). Eye gaze and individual differences consistent with learned attention in associative blocking and highlighting. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *31*, 830–845.

Larrauri, J. A., & Schmajuk, N. A. (2008). Attentional, associative, and configural mechanisms in extinction. *Psychological Review*, *115*(3), 640–675.

Le Pelley, M. E., & McLaren, I. P. L. (2003). Learned associability and associative change in human causal learning. *The Quarterly Journal of Experimental Psychology Section B*, *56*(1), 68–79.

Le Pelley, M. E., Oakeshott, S. M., Wills, A. J., & McLaren, I. P. L. (2005). The outcome specificity of learned predictiveness effects: Parallels between human causal learning and animal conditioning. *Journal of Experimental Psychology: Animal Behavior Processes*, *31*(2), 226–236.

Lewandowsky, S., Roberts, L., & Yang, L. X. (2006). Knowledge partitioning in categorization: Boundary conditions. *Memory & Cognition*, *34*(8), 1676–1688.

Little, D. R., & Lewandowsky, S. (2009). Beyond non-utilization: Irrelevant cues can gate learning in probabilistic categorization. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(2), 530–550.

Livesey, E., & McLaren, I. (2010). An elemental model of associative learning and memory. In E. M. Pothos & A. J. Wills (Eds.), *Formal approaches to categorization* (chap. 7). Cambridge, UK: Cambridge University Press.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN:

A network model of category learning. *Psychological Review*, *111*(2), 309–332.

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

McDonnell, J. V., & Gureckis, T. M. (2010). Adaptive clustering models of categorization. In E. M. Pothos & A. J. Wills (Eds.), *Formal approaches to categorization* (chap. 10). Cambridge, UK: Cambridge University Press.

McLaren, I. P. L., & Mackintosh, N. J. (2000). An elemental model of associative learning: I. Latent inhibition and perceptual learning. *Animal Learning and Behavior*, *28*(3), 211–246.

McLaren, I. P. L., & Mackintosh, N. J. (2002). An elemental model of associative learning: II. generalization and discrimination. *Animal Learning and Behavior*, *30*, 177–200.

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.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*, 207–238.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology*, *115*, 39–57.

Nosofsky, R. M. (2010). The generalized context model: An exemplar model of classification. In E. M. Pothos & A. J. Wills (Eds.), *Formal approaches to categorization* (chap. 2). Cambridge, UK: Cambridge University Press.

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.

Nosofsky, R. M., & Johansen, M. K. (2000). Exemplar-based accounts of "multiple-system" phenomena in perceptual categorization. *Psychonomic Bulletin & Review*, *7*(3), 375–402.

Oswald, C. J. P., Yee, B. K., Rawlins, J. N. P., Bannerman, D. B., Good, M., & Honey, R. C. (2001). Involvement of the entorhinal cortex in a process of attentional modulation: Evidence from a novel variant of an IDS/EDS procedure. *Behavioral neuroscience*, *115*(4), 841–849.

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

Pothos, E. M., & Chater, N. (2002). A simplicity principle in unsupervised human categorization. *Cognitive Science*, *26*, 303–343.

Pothos, E. M., Chater, N., & Hines, P. (2010). The simplicity model of unsupervised categorization. In E. M. Pothos & A. J. Wills (Eds.), *Formal approaches to categorization* (chap. 9). Cambridge, UK: Cambridge University Press.

Rehder, B., & Murphy, G. L. (2003). A knowledge-resonance (KRES) model of category learning. *Psychonomic Bulletin and Review*, *10*(4), 759–784.

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.

Rodrigues, P. M., & Murre, J. M. J. (2007). Rules-plus-exception tasks: A problem for exemplar models? *Psychonomic Bulletin & Review*, *14*, 640–646.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by back-propagating errors. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing* (Vol. 1). Cambridge, MA: MIT Press.

Sanborn, A. N., Griffiths, T. L., & Navarro, D. (2006). A more rational model of categorization. In R. Sun & N. Miyake (Eds.), *Proceedings of the 28th annual conference of the cognitive science society.* Mahwah, NJ: Erlbaum.

Shanks, D. R. (1985). Forward and backward blocking in human contingency judgement. *Quarterly Journal of Experimental Psychology: Comparative & Physiological Psychology*, *37B*, 1–21.

Shepard, R. N., Hovland, C. L., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs*, *75*(13). (Whole No. 517)

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

Tenenbaum, J. B., & Griffiths, T. L. (2003). Theory-based causal inference. In S. Becker, S. Thrun, & K. Obermayer (Eds.), *Advances in neural information processing systems* (Vol. 15, pp. 35–42). Cambridge, MA: MIT Press.

Treat, T. A., Kruschke, J. K., Viken, R. J., & McFall, R. M. (2010). Application of associative learning paradigms to clinically relevant individual differences in cognitive processing. In T. R. Schachtman & S. Reilly (Eds.), *Conditioning and animal learning: Human and non-human animal applications* (pp. **–**). Oxford, UK: Oxford University Press.

Treat, T. A., McFall, R. M., Viken, R. J., & Kruschke, J. K. (2001). Using cognitive science methods to assess the role of social information processing in sexually coercive behavior. *Psychological Assessment*, *13*(4), 549–565.

Treat, T. A., McFall, R. M., Viken, R. J., Kruschke, J. K., Nosofsky, R. M., & Wang, S. S. (2007). Clinical cognitive science: Applying quantitative models of cognitive processing to examine cognitive aspects of psychopathology. In R. W. J. Neufeld (Ed.), *Advances in clinical cognitive science: Formal modeling of processes and symptoms* (pp. 179–205). Washington, D.C.: American Psychological Association.

Vigo, R. (2006). A note on the complexity of Boolean concepts. *Journal of Mathematical Psychology*, *50*, 501–510.

Vigo, R. (2009). Categorical invariance and structural complexity in human concept learning. *Journal of Mathematical Psychology*, *53*, 203–221.

Wills, A. J., Lavric, A., Croft, G. S., & Hodgson, T. L. (2007). Predictive learning, prediction errors, and attention: Evidence from event-related potentials and eye tracking. *Journal of Cognitive Neuroscience*, *19*(5), 843–854.

Yang, L.-X., & Lewandowsky, S. (2003). Context-gated knowledge partitioning in categorization. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *29*(4), 663–679.

Yang, L.-X., & Lewandowsky, S. (2004). Knowledge partitioning in categorization: Constraints on exemplar models. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *30*(5), 1045–1064.