

Learned Attention

John K. Kruschke

*Department of Psychological and Brain Sciences
Indiana University*

Abstract—Unlike many approaches to machine learning, human learning involves selective attention. When confronted by new things to learn, people can rapidly shift attention, thereby increasing speed of acquisition and decreasing interference with previous knowledge. The shift of attention is itself learned, so that attention is allocated to particular cues in particular contexts. While selective attention benefits acquisition, it can also lead to distortions of knowledge that are evident when the knowledge is transferred to novel situations. Several mathematical models have been designed to implement selective attention in learning; the models quantitatively fit human performance in many experiments. This presentation reviews various research projects of the author.

Index Terms—Attention, back propagation, Bayesian learning, connectionist model, eye tracking, incremental learning, individual differences, intradimensional extradimensional shift, learning, mixture of experts, rules and exceptions.

I. HUMAN LEARNING VS MACHINE LEARNING

From a rational engineering perspective, learning algorithms should utilize all the information that is available. By including all the information given, an algorithm can, in principle, correctly learn all the true correlations and contingencies among the myriad cues.

But human learning is highly selective. People pay attention to the information that is (or seems to be) relevant for the task at hand. This selectivity is not merely a capacity limitation. Appropriate attention to selected cues, and suppression of other cues, can accelerate learning and reduce interference with previously acquired knowledge.

While selective attention enhances acquisition of the specific training data, it can also distort the knowledge that is acquired. Such distortions lie unnoticed until there arise novel situations to which the knowledge must be generalized. People exhibit many interesting generalization patterns that few “normative” learning algorithms can predict.

This presentation summarizes some of my research on this topic. This research lives in the context of a wealth of research by other scientists, but space constraints prevent me from including a review of the field. Links to the literature can be found in the original publications, and in a broader review [10]. In the present article, I will first review some benchmark laboratory phenomena that indicate a strong role for attention in human learning. Then I will review some work regarding real-world manifestations of attention in learning. Finally, I will review several mathematical models

of attention in learning, and a general framework in which they fit.

II. LABORATORY PHENOMENA INDICATIVE OF ATTENTION IN LEARNING

I will restrict my discussion to experiments in which people learn what simple response to make in the presence of various cues. The learner is seated at a standard computer and, on a learning trial, is shown a few simple cues on the monitor. The cues might be color patches or geometric figures or words. The learner is prompted to guess which of several keys to press, and after a response is made, the computer reveals the correct answer. Trials continue and the learner’s accuracy improves.

Research on human learning has found many phenomena that are naturally explained by selective attention to cues. One of these findings is that people are faster to learn response mappings for which fewer cues are relevant. A classic demonstration of this was provided by Shepard, Hovland and Jenkins [20]. In their experiments, participants learned to classify stimuli into two categories. The stimuli varied on three binary dimensions. In the prototype structure (which they called Type IV), each of the two categories had a prototype and three other members that were just one feature different from the prototype. In the XOR structure (which they called Type II), the categories were defined by an exclusive-OR on two dimensions, with the third dimension being irrelevant. Despite the fact that the prototype structure has convex categories that are linearly separable, people learn it more slowly than the XOR structure. The relative ease of the XOR structure suggests that people learn to selectively attend to the two relevant dimensions and ignore the third, irrelevant dimension. Models that do not selectively attend to dimensions are quite challenged by the XOR advantage, but models that include selective attention naturally accommodate that result [5], [18].

Another dramatic example involves stimuli that vary on two continuous dimensions (rather than three binary-valued dimensions). The stimuli were rectangles that could have different heights, containing an interior vertical segment that could have different lateral positions. There were eight stimuli to be classified, and their specific heights and line positions defined eight points in the two-dimensional stimulus space. These eight points occupied the corners of an octagon, like a stop sign. In one structure to be learned, the stimuli

at the four corners on the left of the octagon were in one category, and the stimuli at the four corners on the right of the octagon were in the other category. This categorization can be accurately learned by attending only to the horizontal stimulus dimension. That is, the vertical dimension can be filtered (excluded) without loss of accuracy, and so this is called a filtration task. In another structure, the category labels were simply rotated 45° , such that the four corners on the upper right of the octagon were in one category and four corners on the lower left of the octagon were in the other category. This categorization requires that both dimensions be processed for complete accuracy. Because information from both dimensions must be condensed into a single categorical response, this is called a condensation task. People learned the filtration structure much faster than the condensation structure. This filtration advantage is very difficult to mimic by models such as vanilla backpropagation that can arbitrarily re-represent the stimuli. But models that learn to selectively attend to the stimulus dimensions accommodate the human difference very well [6].

The previous two examples emphasized the idea that people can learn to pay attention to stimulus dimensions. People can also learn to pay attention to different aspects of stimuli depending on the particulars of the stimulus. Michael Erickson and I have proposed that people can learn to attend to rule-like representations of stimuli or exemplar-like representations depending on the stimulus. We studied category structures defined on two continuous dimensions, in which most of the stimuli could be classified by a simple threshold-rule on one dimension, but for which there were a few exceptions, located at the periphery of the training domain, that required attention to both dimensions. Many people learned the rule-described items much faster than exemplar-based models can explain [14]. People also extrapolated beyond the training domain according to the rule, even for test probes that were closer to the learned exceptions than to any rule training item. This extrapolation behavior, and crossovers in the learning curves, are captured by a mixture-of-experts [3] model that learns to selectively attend to rules or exemplars [1], [2].

Analogous structures have been applied to function learning, in which the response is not a categorical value but a continuous value. Michael Kalish, Stephen Lewandowsky and I trained people on input-output functions that were linear over most of their domain, with the exception of a few isolated points. When extrapolating beyond the training range, people responded bimodally, either continuing the linear function or responding with a value consistent with the exceptions. These results were modeled with a mixture-of-experts architecture that learns to attend either to linear functions or exemplars [4].

Among my favorite phenomena are two called highlighting and blocking. In these situations, the stimuli are merely

TABLE I
CORE OF BLOCKING AND HIGHLIGHTING EXPERIMENT DESIGNS.

Phase	Design			
	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)		PE.PL→? (L) I→? (E)	

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

present/absent cues. As shown in Table I, the training proceeds in phases, with different cue combinations and outcomes being introduced along the way. In blocking, the first phase has a single cue A indicate the outcome X, denoted A→X. Once that association is well learned, cue A is then accompanied simultaneously by a second cue, B. For purposes of comparison, this second phase also intermixes trials of C.D→Y, which occur as many times as trials of A.B→X. If all that matters for learning is the number times a cue and outcome co-occur, then B should be learned to indicate X just as strongly as D should be learned to indicate Y. But in subsequent tests with the ambiguous cue combination B.D, people prefer to respond Y (not equally X or Y). This result suggests that learning about the B cue has been blocked, i.e., prevented, by previous learning about the A cue; hence the appellation “blocking.” Aspects of blocking can be explained without appealing to selective attention, but there are other aspects (such as learning after blocking, to be described below) that are naturally explained in terms of attention, as follows: When people are learning A.B→X, cue B distracts attention away from the already learned cue A, causing a reduction in the strength of the correct response. This weakness can be corrected by simply directing attention back to A, away from B. Thus, people learn to ignore B and attend to A.

The highlighting procedure is complementary to blocking. Table I shows that an early trained outcome, E, is indicated by an imperfect predictor I and an perfect predictor, PE. Those cases continue in a later phase, along with cases of a new, later-trained outcome: I.PL→L. If people learned the simple underlying symmetry of the cues, they would know that PE and PL are equally good predictors of their respective outcomes. But when tested with the ambiguous probe PE.PL, people prefer response L (not E or L equally). This result is very difficult for many models of learning, but can be naturally explained by shifting attention: When learning I.PL→L, cue I is already associated with the now-incorrect response E, and so people shift attention away from I toward PL. People learn an association from PL to L, but also learn to shift attention to PL, especially in the context of cue I.

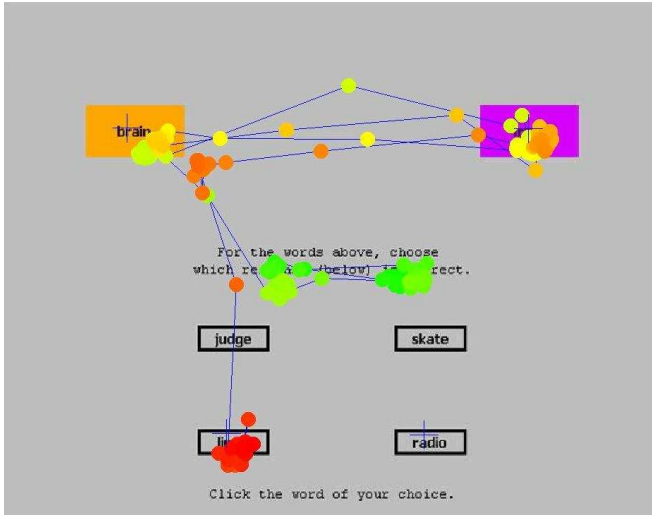


Fig. 1. Example of a stimulus display with an eye-gaze trajectory superimposed.

Recent eye-tracking research reveals that people look longer at the cues that theory predicts should get more attention [16]. Figure 1 shows an example of a stimulus display with an eye gaze trajectory superimposed. In the blocking design, people look longer at cue D than at cue B. In the highlighting design, people look longer at cue PL than at cue PE .

Interestingly, the magnitude of blocking and highlighting covaries across individuals [16]. People who exhibit strong highlighting tend to also exhibit strong blocking. This correlation is predicted by the hypothesis that attentional shifting and learning varies across individuals, and is stronger in some persons than in others. When the attentional parameters of a mathematical model (described below) are varied, the model naturally covaries the magnitudes of blocking and highlighting. Other parameters in the model do not account for the covariation, nor do other models.

Finally, a major finding that indicates learned attention is perseveration on previously relevant cues when the response contingencies are changed. Consider, for example, stimuli that vary in color and shape. Initially, the shape is relevant for the correct response and the color is irrelevant. Then, unannounced to the learner, the corrective feedback changes such that now color is relevant and shape is no longer relevant. This extradimensional shift is more difficult to learn than an intradimensional shift, in which the same dimension remains relevant. I constructed a novel version of this task to challenge theories of learning [7]. Stimuli varied on three binary dimensions, and the initial category structure was an XOR on two dimensions with the third dimension irrelevant. In the shifted structure, just one dimension when relevant, and it was either the previously irrelevant dimension or one of the initially XOR-relevant dimensions. Results showed that

a shift to the previously irrelevant dimension was much more difficult than a shift to a previously relevant dimension, even though neither of the previously relevant dimensions was individually correlated with the correct response.

Analogous studies have been conducted on learning after blocking and after highlighting [11], [13]. Learning about a previously blocked cue is retarded compared to a control cue, and learning about a previously highlighted cue is improved relative to a control cue.

In summary, laboratory research has established many results consistent with the hypothesis the people selectively attend during learning. Structures with fewer relevant dimensions are easier to learn than structures with many relevant dimensions. People extrapolate according to rules even when the nearest trained items are exceptions. People generalize from learned cues as if some cues were selectively blocked or highlighted during learning. People look longer at some cues than other. When contingencies change such that new cues become relevant, the ease of learning the shift depends on what was previously learned to be relevant. Many of these findings are difficult for learning algorithms that attend to all the information all the time, but the phenomena appear naturally from models that learn to selectively attend.

III. REAL WORLD MANIFESTATIONS OF ATTENTION IN LEARNING

Laboratory experiments are intentionally constructed to minimize the influence of extraneous factors such as background knowledge and differential cue salience. This disengagement from accidental influences is standard scientific procedure. If we are interested in the properties of gravity, we study the behavior of a sphere in a vacuum; we do not study the behavior of a leaf in a thunderstorm. But people learn in the thunderstorm of real life, not only the vacuum of the laboratory. Does real-world learning also involve learned selective attention?

One everyday situation is learning from the results of a web search. The search engine returns a list of web pages, and users typically explore the list beginning with the links at the top. Thus, the ordered list biases users to learn about some items before others. This ordering is analogous to the ordered phases of training in blocking and highlighting (recall Table I). In unpublished research done in 2001 with collaboration of an undergraduate honors student named Nancy Aleman, we told participants that they were to learn about the qualities of whitewater rafts. This knowledge could be used for decisions about which rafts to rent or purchase. Figure 2 shows an example of a web page seen by the learners. The page gives a description of a raft, with particular features pointed out along with a quality rating provided by a fictitious independent group, the Society of Whitewater Rafters.

In one version of the experiment, learners browsed a couple dozen pages to learn about the rafts currently available on

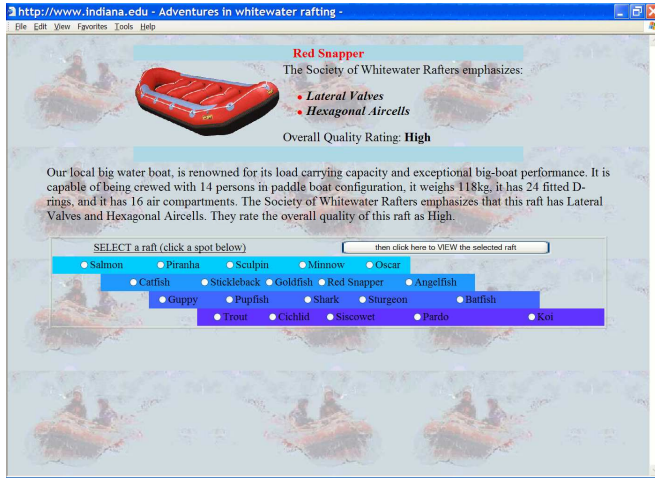


Fig. 2. Example of a web page for learning about white water rafts.

the market. Then they viewed a few pages that purported to show prototypes of rafts that manufacturers were considering bringing to market. Participants predicted the quality of each raft based on the features of the raft. Results showed a strong highlighting effect in predicted quality: Rafts with the imperfectly predictive feature (I) were given the earlier-learned quality (E), and rafts with a combination of the two perfectly predictive features (PE and PL) were given the later-later quality (L).

Real world learning often involves situations in which cues are only probabilistic indicators of outcomes. This uncertainty raises many questions as we go through an ordinary day. Does that cloud formation signal rain later today? Does that shade of green on the rind indicate the melon is ripe? Does that person's smile indicate friendship or just politeness? Real world learning also involves cues of differential salience, which might have little to do their validity. Advertisers take advantage of this fact, attracting our attention to their products using salient cues that have no relevance to the product itself.

Laboratory studies have confirmed that there are attentional trade-offs between salient and valid cues, especially in probabilistic settings (e.g., [15]). Individual differences in perceived salience can also have dramatic effects on learning. Teresa Treat and others [21] examined how men perceive and learn about aspects of women in photographs. The female models in the photographs, many from mass-market magazines, varied in facial affect (happy, neutral or sad) and skin exposure (e.g., strapless top or long sleeve blouse). Other assessments measured individual differences in the subjective salience of affect and exposure. That is, different male subjects arrived at the experiment with different tendencies pay attention to affect or exposure. In the learning tasks, affect was the relevant dimension for some category structures, and exposure was the relevant dimension

for other structures. Results showed that the individual's subjective salience influenced the difficulty of learning about each dimension, and, moreover, that it may have been difficult to change how much attention was allocated to each each dimension. Thus, these dimensions of affect and exposure might be relatively integral or inseparable, so that attention cannot be selectively allocated to them.

There are potentially important ramifications of this type of research on learning and shifting attention to everyday cues. Some people may have dysfunctional attention to cues; e.g., some men might pathologically ignore a woman's affect, some women might pathologically attend to their own body size [22], some drug abusers might pathologically attend to drug-relevant cues. Part of therapy may be retraining attention or retraining what the cues are associated with.

IV. MATHEMATICAL MODELS OF ATTENTION IN LEARNING

Once the detailed empirical regularities are established, a major goal of science is to build formal models of those phenomena. One benefit of formal modeling is that the syntax of the formalism generates predictions of the theory. Without that independent engine for publicly producing predictions, it is only the theorist's idiosyncratic intuition that mysteriously conjures predictions. Another benefit of formal models is that detailed quantitative predictions can be generated, instead of imprecise qualitative predictions.

Much of my effort has been devoted to creating models of attention in learning. Figure 3 shows the general framework of my approach, along with several specific instantiations of that framework. The general framework (upper left panel of Figure 3) assumes that whenever a mapping from cues to outcomes is being learned, there are actually two sub-mappings being created. First, the cues are mapped to an allocation of attention over the cues. The diagram indicates attentional gates as triangles pointing at each cue. The second mapping connects attentionally filtered cues to outcomes. The target for the attentional allocation is determined by a rapid shift of attention that occurs before the mappings are learned. The attention is shifted to reduce error, or, equivalently, to increase the probability of generating the correct response. Once the attention has been shifted, then the mappings are adjusted to better generate the target attention and target outcome.

The general framework makes no commitments regarding the nature of the learned mappings. The mappings could be instantiated as a linear weighting, as several layers in a backpropagation network, as a layer of exemplars with weighted influence, as a set of productions or condition-consequent rules, as a set of probabilistic hypotheses with degrees of belief, etc.

Over the years I have created different instantiations of the general framework. The different instantiations were

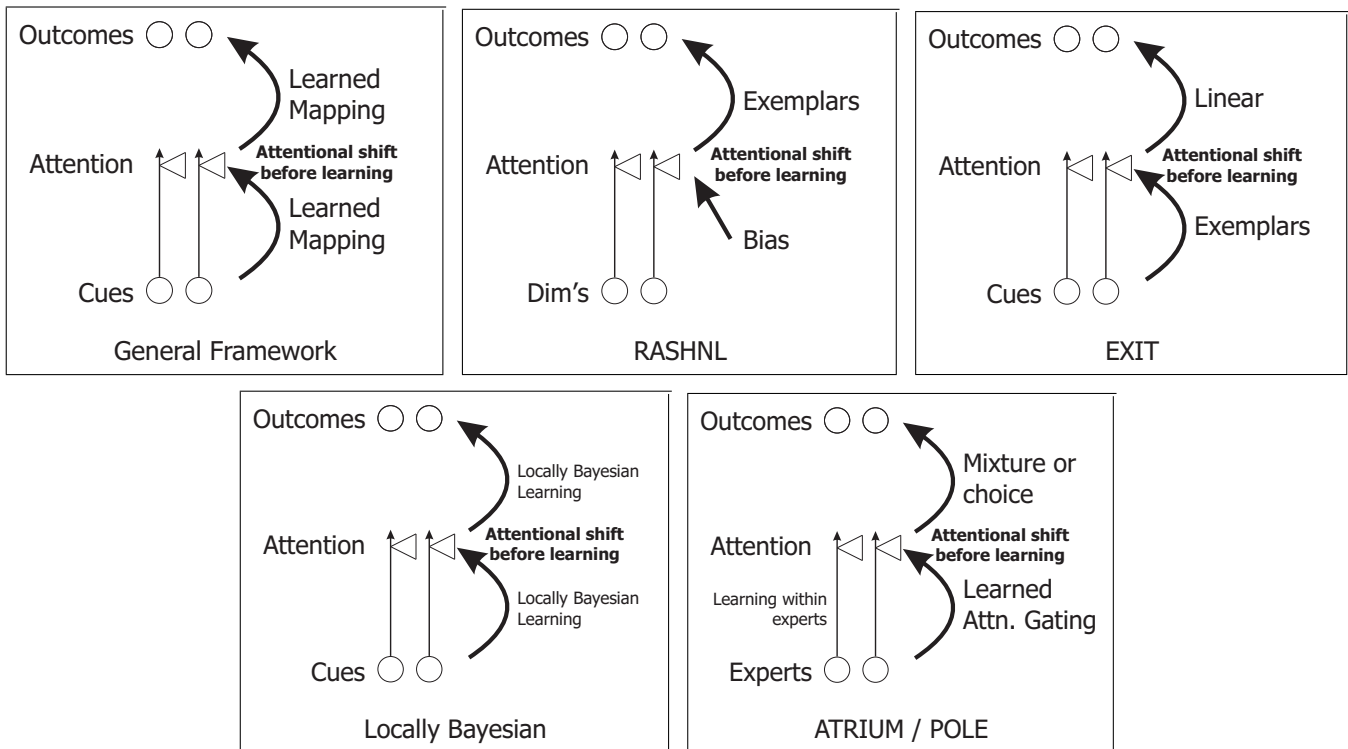


Fig. 3. Upper left panel shows general framework for models of attention shifting and learning. Upper middle: RASHNL model [15]. Upper right: EXIT model [8], [9]. Lower left: Locally Bayesian learning [12]. Lower right: Mixture of experts in ATRIUM [1], [2] or POLE [4].

motivated by different antecedent models in the literature, and by different types of stimuli in different empirical procedures. In the RASHNL model [15] (upper middle panel of Figure 3), each input node represents a stimulus dimension, not just a value on a dimension. The model has a limited ability to learn attentional distributions however, because it learns just one distribution of attention that is applied universally to all stimuli. Its mapping from attentionally filtered dimensions to outcomes is completely flexible, however, because it is mediated by exemplars (one per unique stimulus). These particular characteristics of the model were motivated as incremental variations on the model's ancestors, namely ALCOVE [5] and the generalized context model [17]. The RASHNL model has been applied to data from experiments with probabilistic outcomes and dimensions of differing salience [15], [21].

In the EXIT model (upper right panel of Figure 3), each input node represents a present/absent cue. The mapping from cues to attention distribution is very flexible, being mediated by exemplars, i.e., configurations of cues. This allows the model learn exemplar-specific allocations of attention, instead of universal attentional allocation as in the RASHNL model. The mapping from attentionally filtered cues to outcomes is merely linear, however, because all the applications of EXIT so far have not demanded non-linear mappings. Another reason for the linear associations to the outcome nodes is

that EXIT's ancestry includes the Rescorla-Wagner model [19], which is linear. EXIT has been applied to the blocking and highlighting effects with great success [8], [9], [11], [16].

A generalization of RASHNL and EXIT could include exemplar-mediated mappings in both layers. This enhancement would allow the resulting model to learn arbitrarily complex mappings from cues to attention allocations, and arbitrarily complex mappings from attentionally filtered cues to outcomes. The cost of such an embellishment is an increase in parameters, but the benefit might be that a single model could capture an even broader spectrum of phenomena.

The lower right panel of Figure 3 indicates that the mixture-of-experts approach is also a variation of the general framework. What gets attended to, however, is not single cues or dimensions, but expert modules that represent or process the stimuli in distinct ways. When corrective feedback is supplied on a learning trial, the model allocates attention to the expert module that best predicts the correct outcome. Then the model learns that attentional allocation, and the selected module adjusts its internal mapping. This approach has been successfully applied to data from learning of rules and exceptions in categorization [1], [2], [14] and function learning [4].

My most recent endeavors [12] have considered Bayesian learning of the mappings. In a Bayesian approach, the model

begins with a set of hypothesized candidate mappings, and a prior degree of belief in each candidate mapping. On each training trial, the model shifts its belief away from hypotheses that are inconsistent with the outcome on that trial. Belief is loaded onto hypotheses that best account for the training items. The proper treatment of this situation would consider all possible cross combinations of the two layers of mappings. In this proper treatment, learning would shift belief to the *combined* mapping that best captures the training items. In my locally Bayesian approach, however, the model does not have the luxury of considering the full Cartesian product of all hypotheses. Instead, each layer can only shift belief to hypotheses that account for the input and target on that particular layer. To achieve this, each layer gets a target that is propagated backwards from the outcome target. At the attention layer, the target for the attention mapping is the distribution of attention that maximizes the probability of outcome target. This target constitutes a rapid shift of attention before Bayesian learning occurs locally on the attention hypothesis space. The backpropagated target also depends on the beliefs held by the outcome layer at the time. Because of this dependency, the locally Bayesian model can be said to first adjust the data (its internal target) to fit its beliefs, before it adjusts its beliefs to fit the data.

The lower left panel of Figure 3 suggests that this locally Bayesian process fits within the general framework I have expounded. Locally Bayesian learning achieves some behavior that cannot be explained by the other instantiations, however. For example, Bayesian updating naturally explains “backward blocking,” which is analogous to the blocking effect (see Table I) but with the two training phases reversed. Various other Bayesian approaches in the literature cannot account for highlighting, which requires rapid attentional shifting and learning as in the locally Bayesian approach.

These various models have been quantitatively fit to a variety of human performance data. The models account for human behaviors that are very challenging for other models that do not have selective attention. What all the instantiations of Figure 3 share is a commitment to the general principle that people can, at least for some stimuli and some situations, rapidly shift attention to different sources of information, in order to learn the desired response quickly and consistently with previously acquired knowledge.

ACKNOWLEDGMENT

I thank the session organizer, Dr. Zhengyou Zhang, for inviting me to present this summary. Address correspondence to John K. Kruschke, Department of Psychological and Brain Sciences, Indiana University, 1101 E. 10th St., Bloomington IN 47405 USA. Electronic mail may be sent to kruschke@indiana.edu. More information is available at <http://www.indiana.edu/~kruschke/>.

REFERENCES

- [1] M. A. Erickson and J. K. Kruschke. Rules and exemplars in category learning. *Journal of Experimental Psychology: General*, 127(2):107–140, 1998.
- [2] M. A. Erickson and J. K. Kruschke. Rule-based extrapolation in perceptual categorization. *Psychonomic Bulletin & Review*, 9(1):160–168, 2002.
- [3] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3:79–87, 1991.
- [4] M. L. Kalish, S. Lewandowsky, and J. K. Kruschke. Population of linear experts: Knowledge partitioning and function learning. *Psychological Review*, 111(4):1072–1099, 2004.
- [5] J. K. Kruschke. ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, 99:22–44, 1992.
- [6] J. K. Kruschke. Human category learning: Implications for backpropagation models. *Connection Science*, 5:3–36, 1993.
- [7] J. K. Kruschke. Dimensional relevance shifts in category learning. *Connection Science*, 8:201–223, 1996.
- [8] J. K. Kruschke. The inverse base rate effect is not explained by eliminative inference. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 27:1385–1400, 2001.
- [9] J. K. Kruschke. Toward a unified model of attention in associative learning. *Journal of Mathematical Psychology*, 45:812–863, 2001.
- [10] J. K. Kruschke. Category learning. In K. Lamberts and R. L. Goldstone, editors, *The Handbook of Cognition*, chapter 7, pages 183–201. Sage, London, 2005.
- [11] J. K. Kruschke. Learning involves attention. In G. Houghton, editor, *Connectionist Models in Cognitive Psychology*. Psychology Press, London, 2005.
- [12] J. K. Kruschke. Locally Bayesian learning with applications to retrospective reevaluation and highlighting. *Psychological Review*, in press.
- [13] J. K. Kruschke and N. J. Blair. Blocking and backward blocking involve learned inattention. *Psychonomic Bulletin & Review*, 7:636–645, 2000.
- [14] J. K. Kruschke and M. A. Erickson. 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*, pages 514–519, Hillsdale, NJ, 1994. Erlbaum.
- [15] J. K. Kruschke and M. K. Johansen. A model of probabilistic category learning. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 25(5):1083–1119, 1999.
- [16] J. K. Kruschke, E. S. Kappenman, and W. P. Hetrick. 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, 2005.
- [17] R. M. Nosofsky. Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology*, 115:39–57, 1986.
- [18] R. M. Nosofsky, M. A. Gluck, T. J. Palmeri, S. C. McKinley, and P. Glauthier. Comparing models of rule-based classification learning: A replication of Shepard, Hovland, and Jenkins (1961). *Memory & Cognition*, 22:352–369, 1994.
- [19] R. A. Rescorla and A. R. Wagner. A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and non-reinforcement. In A. H. Black and W. F. Prokasy, editors, *Classical conditioning II: Current research and theory*, pages 64–99. Appleton-Century-Crofts, New York, 1972.
- [20] R. N. Shepard, C. L. Hovland, and H. M. Jenkins. Learning and memorization of classifications. *Psychological Monographs*, 75(13), 1961. Whole No. 517.
- [21] T. A. Treat, R. M. McFall, R. J. Viken, and J. K. Kruschke. Using cognitive science methods to assess the role of social information processing in sexually coercive behavior. *Psychological Assessment*, 13(4):549–565, 2001.
- [22] T. A. Treat, R. M. McFall, R. J. Viken, R. M. Nosofsky, D. B. MacKay, and J. K. Kruschke. Assessing clinically relevant perceptual organization with multidimensional scaling techniques. *Psychological Assessment*, 14(3):239–252, 2002.