AMERICAN PSYCHOLOGICAL ASSOCIATION

Psychological Review

ISSN: 0033-295X

2022, Vol. 129, No. 5, 1104–1143
https://doi.org/10.1037/rev0000381

As Within, so Without, as Above, so Below: Common Mechanisms Can Support Between- and Within-Trial Category Learning Dynamics

Emily R. Weichart, Matthew Galdo, Vladimir M. Sloutsky, and Brandon M. Turner
Department of Psychology, The Ohio State University

Two fundamental difficulties when learning novel categories are deciding (a) what information is relevant and (b) when to use that information. Although previous theories have specified how observers learn to attend to relevant dimensions over time, those theories have largely remained silent about how attention should be allocated on a within-trial basis, which dimensions of information should be sampled, and how the temporal order of information sampling influences learning. Here, we use the adaptive attention representation model (AARM) to demonstrate that a common set of mechanisms can be used to specify: (a) How the distribution of attention is updated between trials over the course of learning and (b) how attention dynamically shifts among dimensions within a trial. We validate our proposed set of mechanisms by comparing AARM's predictions to observed behavior in four case studies, which collectively encompass different theoretical aspects of selective attention. We use both eye-tracking and choice response data to provide a stringent test of how attention and decision processes dynamically interact during category learning. Specifically, how does attention to selected stimulus dimensions gives rise to decision dynamics, and in turn, how do decision dynamics influence which dimensions are attended to via gaze fixations?

Keywords: categorization, learning, decision dynamics, eye tracking

When asked to describe an object, we instinctively do so in terms of its components, or *dimensions*. To describe a jacket, we might note dimensions like its color or size, where its pockets are placed, or any insignia it has. When assigning objects to different categories, certain dimensions are often more relevant than others depending on the demands of the task. Distinguishing between spring and winter jackets, for example, might require us to specifically note dimensions like material, thickness, and types of closures, whereas distinguishing between formal and casual jackets might depend on dimensions like length and style.

How do we figure out which dimensions are relevant to a particular task, and how do we use that information to categorize new items? Theoretical accounts of category learning have suggested that over the course of experience with many items, humans gradually build up associations between features (i.e., "linen" and "wool" could be considered to be features of the "material" dimension) and the available category labels (i.e., spring and winter jackets). As more pairings between stimuli and category labels are presented, the observer learns that a subset of dimensions is particularly relevant for identifying category membership among all sources of information that are available.

Several models have described learning as a process of selectively attending to the most category-diagnostic dimensions to support an increase in accuracy across trials (e.g., Kruschke, 1992; Love et al., 2004; R. Nosofsky, 1986). Although attention is often described as a latent mechanism, the general mode of learning via selective attention has garnered theoretical support from eye-tracking work. Results consistently show an increase in the proportions of fixations to task-relevant dimensions, which co-occur with increasing categorization accuracy (McColeman et al., 2014; Rehder & Hoffman, 2005a, 2005b). Despite these findings, the impact of learning on subsequent, generalized behaviors of information sampling and decision-making has remained underexplored. In other words, how does the knowledge we acquire through learning, such as memories of previous items and the task relevance of individual dimensions, impact the manner in which we seek out information about new stimuli? As suggested by Rehder and Hoffman (2005a, 2005b), one might reasonably assume that dimensions are fixated during each trial in proportion to their respective attention weights. Intriguing experimental work by Blair and colleagues (Blair et al., 2009; Chen et al., 2013; McColeman et al., 2014; Meier & Blair, 2013), however, has indicated that there might be more to the story.

This article was published Online First July 18, 2022.
Emily R. Weichart https://orcid.org/0000-0001-5535-5704
Matthew Galdo https://orcid.org/0000-0002-1279-3859
Brandon M. Turner https://orcid.org/0000-0003-0966-9301
This work was supported by Faculty Early Career Development Program Grant CAREER1847603 awarded to Brandon M. Turner by the National Science Foundation (NSF), and Grant RO1HD078545 awarded to Vladimir M. Sloutsky by the National Institutes of Health (NIH).

Preliminary results were first presented at the Annual Meeting of the Mathematical Psychological Society on July 1, 2021 (Emily R. Weichart). An early version of the article was posted on PsyArXiv on June 29, 2021 and is accessible at http://doi.org/10.31234/osf.io/94csh. This study was not preregistered. Model code will be made available upon publication at https://github.com/MbCN-lab. The data used in Case Study 1 will be available upon reasonable request. The data used in Case Study 2 were made freely available online by Meier and Blair (2013) at https://doi.org/10.1016/j.cognition.2012.09.014.

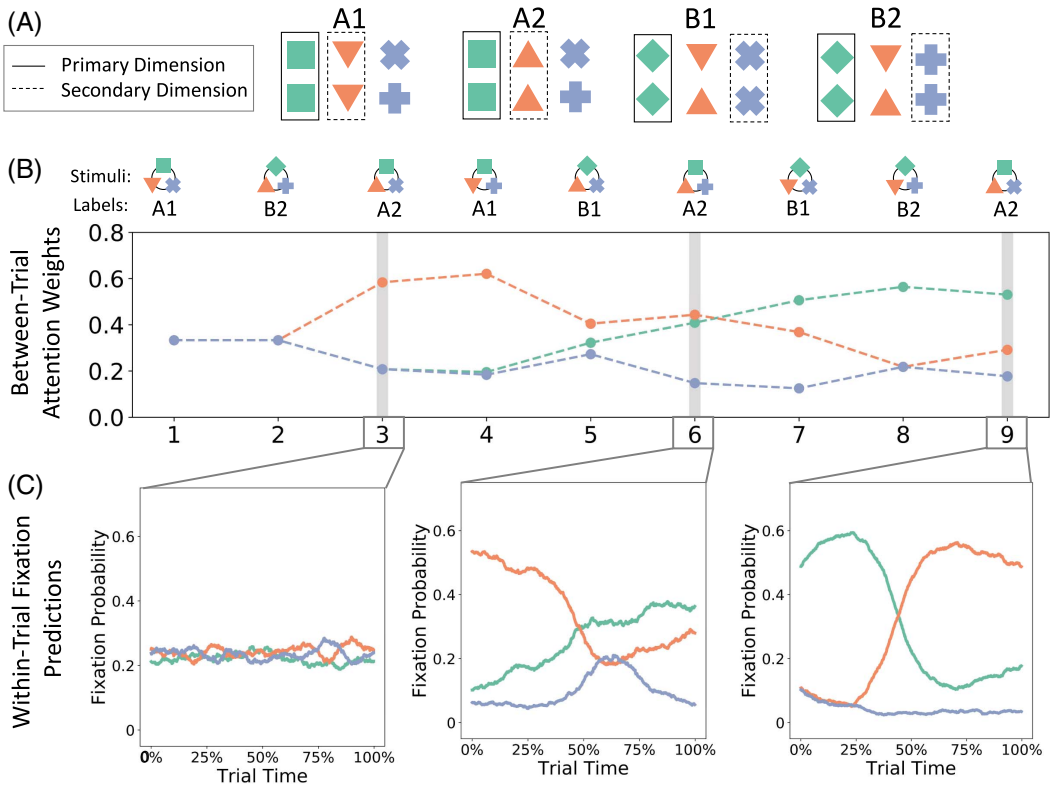
Correspondence concerning this article should be addressed to Brandon M. Turner, Department of Psychology, The Ohio State University, 225 Psychology Building, 1835 Neil Avenue, Columbus, OH 43201, United States. Email: turner.826@gmail.com

In the paradigm illustrated in Figure 1A, stimuli were constructed using a hierarchical category structure where one superordinate dimension (i.e., rotation of the green square) indicated which of two subordinate dimensions was relevant to each trial (i.e., rotation of the orange triangle or the purple cross). While fixations were evenly distributed across dimensions early in the task, participants soon learned to consistently orient to the superordinate dimension as each new trial was presented (Figure 1B, C). Importantly, participants subsequently fixated to one subordinate dimension and ignored the other, depending on the feature identity within the superordinate dimension. In other words, participants tended to only fixate to the two dimensions that were relevant to each trial before making a response, despite all dimensions being equally predictive of category membership on average. These results indicate that humans not only prioritize the most relevant dimensions to make accurate categorization decisions, but also make ongoing decisions within-trial about which sources of information to sample next and when to terminate the sampling process with a response.

The goal of the current article is to establish a common set of mechanisms for allocating attention to relevant dimensions *between trials* over the course of learning, and sampling sources of information *within trials* over the course of individual decisions. We focus on the adaptive attention representation model (AARM), which was described and validated using data from five benchmark category learning paradigms in our previous work (Galdo et al., 2021). AARM inherits its conceptual basis from context theory, which suggests that the feature and category information associated with previously-experienced items are stored in memory as discrete episodic traces (Medin & Schaffer, 1978). As a dynamic extension to the generalized context model (GCM; R. Nosofsky, 1986), AARM describes how category representations are formed according to the similarity between new stimuli and stored exemplars, and are influenced by attention. The amount of attention allocated to each dimension is updated according to trial-level feedback, in a manner that is intended to optimize future responses with respect to the learner's goals.

Figure 1

Within- and Between-Trial Dynamics



Note. AARM = adaptive attention representation model. (A) Illustration of a hierarchical stimulus structure. Feature values (i.e., 0° or 45° rotation) in the superordinate dimension (green squares) indicated which of the two subordinate dimensions (orange triangles or purple crosses) were relevant for identifying category membership. (B) Attention weights generated by AARM's between-trial module, given the sequence of stimuli shown in the top row. Weights were normalized for illustration. Line colors correspond to the colors of the stimulus dimensions. (C) 100 sequences of dimension fixations were generated using the within-trial module. Plots show mean fixation probabilities to each dimension as a function of the percentage of time within-trial, between stimulus onset and self-termination. Within-trial attention weights were initialized according to the outputs of the between-trial module for the relevant stimulus. See the online article for the color version of this figure.

One major innovation of AARM is that it can be fit to both choice and eye-tracking data simultaneously, such that model-generated attention weights are informed by observed proportions of fixations to each dimension. With these constraints in place, Galdo et al. (2021) demonstrated that AARM could predict increasing proportions of fixations to task-relevant dimensions that co-occurred with increasing accuracy across paradigms of varying complexity (e.g., McColeman et al., 2014; Shepard et al., 1961). Like similar adaptive attention models of category learning (Attention Learning COVEring map [ALCOVE]; Kruschke, 1992; Supervised and Unsupervised STratified Adaptive Incremental Network [SUSTAIN]; Love et al., 2004), however, trial-level attention updates in AARM occur only *after* feedback has been observed. While attention weights on Trial i may covary with proportions of fixations on Trial $i + 1$ on average, the standard model lacks the specificity required to predict stimulus-dependent effects of information sampling like those observed by Blair et al. (2009). Here, we therefore extend the mechanisms of AARM that were presented by Galdo et al. (2021) to explain how humans leverage their experiences to construct a representation of a new stimulus.

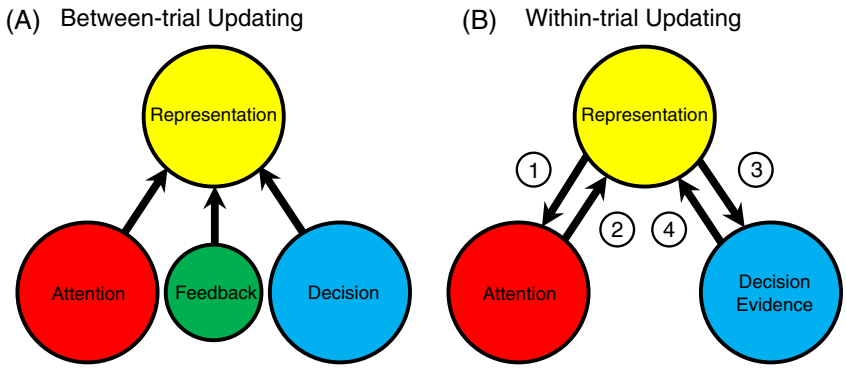
As illustrated in Figure 2, the current work presents the AARM framework as two interrelated modules: (a) a between-trial module to account for feedback-mediated changes in accuracy and attention, and (b) a within-trial module to account for information sampling and decision dynamics. Using insights from accumulation-to-bound decision models (e.g., Ratcliff, 1978) and theoretical notions of pattern completion (Estes, 1994), the within-trial module of AARM makes predictions about how participants decide which dimensions of information to sample (i.e., via fixations), when to sample them, and when to make a response. Taking both modules of AARM together, the current article provides a comprehensive theoretical and computational framework for explaining how knowledge acquisition is fundamentally shaped by the experiences of the learner. Before introducing the mathematical details of AARM, we will first introduce four assumptions that are central to our approach.

Attention Is the Mechanism of Learning

Categorization tasks provide a unique opportunity to study the relationship between learning and attention. From work with animals (Hall, 1991; Le Pelley, 2004) and humans (Bonardi et al., 2005; Kruschke, 1996) demonstrating that learned dimension relevance influences how future stimuli are represented, we gain insight into how attention changes over the course of a task. In a standard type of categorization paradigm, stimuli are designed from a common set of dimensions, each of which can take on one of a unique set of possible feature values. In experiments conducted by Kruschke (1996), for example, stimuli were line drawings of box cars consisting of three dimensions, each of which could take on two possible features: height (tall or short), door position (left or right), and wheel color (black or white). Participants were asked to assign stimuli to arbitrary categories (e.g., Categories "A" and "B") without receiving explicit instructions about how each category was defined. Instead, participants learned the experimentally defined feature-to-category mapping through trial and error with corrective feedback, and learning was assessed through changes in accuracy over multiple trials.

For the sake of illustration, consider an example in which tall box cars belong to Category A, and short box cars belong to Category B. Assuming features are counterbalanced across dimensions, the only way a participant can achieve perfect accuracy is by categorizing stimuli according to the "height" dimension. Although a participant can categorize stimuli on the basis of another dimension like wheel color and be correct on a subset of trials by chance, humans do indeed achieve ceiling-level accuracy in these types of tasks when given sufficient training. In addition to simple "component" mappings (e.g., tall box cars belong to Category A; short box cars belong to Category B) humans can learn more complex "compound" mappings as well (XOR categories; e.g., short, black-wheeled, and tall, white-wheeled box cars belong to Category A; tall, black-wheeled, and short, white-wheeled box cars belong to

Figure 2
Within- and Between-Trial Modules of AARM



Note. AARM = adaptive attention representation model. (A) Between-trial updates to the category representation occur via influences of attention and decision components from the previous trial. (B) Within-trial updates require dynamic interactions among representation, attention, and decision components. First, the representation guides attention to a relevant dimension (1). Attention drives an encoding process for a fixated feature (2) to then update the amount of evidence (3) for each of a set of category responses. The representation is consulted (4) to guide subsequent attentional deployment. See the online article for the color version of this figure.

Category B; Shepard et al., 1961). In general, findings across category learning studies have indicated that human learners (a) gradually acquire knowledge about which dimensions are relevant to the task and (b) make categorization decisions according to which dimensions are perceived to be most relevant (see Ashby & Maddox, 2005; Markman & Ross, 2003, for review).

Categorization models often explain learning as a gradual shift in how stimuli are represented in psychological space. The influential GCM and its descendants have described successful categorization as an outcome of "stretching" multidimensional stimulus representations along relevant dimensions and "shrinking" them along irrelevant dimensions (Kruschke, 1992; Lamberts, 2000; R. Nosofsky, 1986; R. Nosofsky & Palmeri, 1997). As such, stimuli that differ along the relevant dimensions will be perceived as being more dissimilar to one another (i.e., belonging to different categories) than items that differ along the irrelevant dimensions. This manipulation of the psychological object representation comprises the definition of attention among many category learning models, such that allocating attention to a particular dimension distorts the representation across trials accordingly.

The typical use of GCM in explaining attentional phenomena has been to freely estimate attention weights independently across different blocks of an experiment, but the model does not specify a mechanism through which learning occurs. Instead, attention is allocated based on the properties of the category structure, and learning is retrospectively inferred. In the current article, we use intuitions from GCM to outline specific hypotheses about how learning and attention interact, suggesting that *attention itself* is the mechanism for learning.

The between-trial module of AARM uses gradient-based mechanisms to update attention upon observation of category feedback. Because the attention vector weights the influence of plausible feature-to-category mappings when the observer assigns an item to a category, gradient-based updating serves to reallocate attention on every trial in a manner that reduces the likelihood of future errors. As mentioned in the introductory section, our previous work demonstrated that AARM's combination of iterative exemplar storage and attention updating were sufficient for predicting learning-related behaviors across paradigms of varying difficulty (Galdo et al., 2021; Shepard et al., 1961). Here, we additionally describe attention as the mechanism by which information is sampled from individual stimuli, such that fixations at each within-trial timestep are calculated directly from the model's predicted distribution of attention.

AARM's specification of attention as the mechanism for learning departs conceptually from alternative rule-based (RB) classification and Bayesian updating accounts. RB classification models seek to identify the boundary between categories, such that the category label can be determined through a conditional relation or weighted combination of feature values within the current stimulus (Goldberg & Jerrum, 1995; Vapnik, 1998). The Bayesian approach is to construct an internal model of each category through iterative belief updating, and assume that a latent category variable is responsible for generating a distribution of feature values (Anderson, 1991a; Oaksford & Chater, 1998; Tenenbaum & Griffiths, 2001). The sampling emergent attention model (SEA; Braunlich & Love, 2021) combines intuitions from RB and Bayesian learning to account for both information sampling and learning behaviors in the context of categorization problems. Like AARM, SEA consists of two interrelated parts: (a) a concept-learning component, which sorts stimuli into clusters (i.e., Anderson, 1991a) and determines the probability that a new item belongs to each one; and (b) a utility-sensitive sampling component, which performs preposterior analysis to balance the expected information gain of each dimension against a prespecified cost of additional sampling.

Because SEA provides a similarly comprehensive account of within-trial dynamics, we will refer to it throughout the article to provide theoretical contrast to AARM. In particular, we describe SEA as a "rational" alternative to AARM's "mechanistic" approach. As described by Sakamoto et al. (2008), rational theories assume that humans learn to behave optimally within the constraints of the environment. Mechanistic theories, by contrast, aim to predict behavior by defining how information is processed and represented in the brain. For example, parameters representing costs in SEA are primarily used to instantiate different goals (e.g., responding accurately vs. responding quickly), but also comprise the time and effort involved in the perceptual encoding and processing of a stimulus feature. As such, if the observer elects to sample information from a dimension as a result of preposterior analysis, then the relevant feature value is automatically used to update the observer's state of belief about the appropriate category label. Feature encoding in SEA is, therefore, considered to be rational because it uses all known information about the task environment to select the action that will maximize gain and minimize loss: sample information, or make a choice. AARM's within-trial module instead samples information from the dimension with the largest attention weight at each timestep. Attention weights are updated continuously throughout the trial, relative to an evolving working representation of the stimulus. Using familiar terms from the visual search literature (see Itti & Koch, 2001, for review), overt attention (i.e., describing the movement of the eyes) in AARM is explicitly linked to endogenous covert attention (i.e., reflecting latent, goal-directed processing). Encoding a feature value occurs as a function of the cumulative covert attention that is applied to an overtly attended spatial location. We consider feature encoding in AARM to be mechanistic by Sakamoto et al.'s definition because it occurs as a direct consequence of latent theoretical subprocesses. Whereas rational approaches are often considered to have an advantage of precision in terms of the predicted behavior and justification (Anderson, 1991b), mechanistic models are more appropriate for generating novel predictions and understanding nuanced behaviors (Sakamoto et al., 2008). Given the relative merits of each, we use this distinction to highlight how AARM's mechanisms give rise to detailed predictions in various novel contexts, as well as potential departures from the predictions of SEA.

Attention Is Not a Zero-Sum Game

Since seminal work by Sutherland and Mackintosh (1971), attention has often been understood as a fixed-quantity resource that observers use until its limit is reached. The authors presented the inverse hypothesis of animal learning, which described stimulus dimensions in terms of attention units that were modulated by reinforcement (e.g., food reward for correct category discrimination; Mackintosh & Little, 1969). Importantly, the theory imposed the

constraint that attention activation across all dimensions must sum to a constant value, such that increasing the strength of one unit will decrease the strength of the others. Follow-up empirical and theoretical work by Mackintosh (1975), however, rejected the inverse hypothesis in light of evidence that attending to one dimension did not prevent learning of a second dimension in complex stimuli. Nevertheless, the convention of treating attention as a "zero-sum game" persists across many contemporary category learning models, such that attention weights across dimensions are constrained to sum to a constant of one (chosen arbitrarily by, e.g., GCM). Similar intuitions about attention being represented as a constant sum have appeared in perceptual work as well; for example, the assumption that attending to a target stimulus in an array requires equivalent inhibition of distractors (White et al., 2011).

While we do not contest that attentional capacity limitations exist (Brydges et al., 2012; Janssens et al., 2018; Muller et al., 2007; Muller & von Muhlenen, 2000), there is little empirical evidence to suggest that the reserve of attention remains fixed across trials and tasks such that a sum-to-constant constraint is justified. Instead, an expansive literature has shown that task difficulty, perceptual load, and parallel processing affect the extent to which the capacity of the attention system becomes a limiting factor (see Chun et al., 2011, for review). For example, Lavie et al. (Lavie, 1995; Lavie & Cox, 1997; Lavie & Tsal, 1994) have shown that both relevant and irrelevant items are processed in visual search tasks when perceptual load is low, and inhibition of task-irrelevant items only occurs when perceptual load is sufficiently high. The sum-to-constant constraint, however, implies that the capacity limit is reached across tasks, regardless of difficulty.

Other studies have noted fluctuations in attention related to aspects of the stimuli themselves, including perceptual and emotional salience (Theeuwes, 1992, 2010), novelty (Johnston & Schwarting, 1997), and motion (Anderson et al., 2011; Yantis & Egeth, 1999). For example, visual search work showed that the presence of high-salience, task-irrelevant cues significantly impaired subsequent overt attention to task-relevant targets relative to low-salience cues (Baker et al., 2021; Most et al., 2005). One interpretation of the results is that a greater quantity of covert attention continued to be allocated to the high-salience cues despite being removed from the screen before the target even appeared. Considering findings of flexible attention together, it is potentially overly constraining to assume that all attention is known and is entirely allocated to the stimuli intended by a given experimental manipulation, as would be required for inhibition to occur in the presence of a sum-to-constant constraint.

In line with connectionist models such as ALCOVE (Kruschke, 1992) and SUSTAIN (Love et al., 2004) which will be reviewed in detail below, AARM does not adhere to a sum-to-one constraint. Instead, attention to each dimension can fluctuate within and between trials depending on a learned history of predictive reliability, and the sum reserve of available attention is unconstrained. In previous work, Galdo et al. (2021) used model-fitting and comparison methods to evaluate various forms of attentional constraints during category learning. In addition to the standard sum-to-constant constraint, the authors implemented the following within AARM's basic between-trial structure: (a) A norm-to-constant constraint, which allows for different forms of competition between dimensions in addition to the assumption of fixed-quantity attention (e.g., Extended adIT [EXIT]; Kruschke, 2001; Paskewitz & Jones, 2020); (b) least absolute shrinkage and selection operator (LASSO) regularization, which limits the number of dimensions that can be attended within a trial (Park & Casella, 2008); and (c) Ridge regularization, which imposes an upper bound on attention to individual dimensions (Busemeyer Townsend, 1993). The results provided evidence against fixed-quantity attention constraints across five studies, with the model variant containing LASSO regularization and between-dimension competition performing the best overall. These results are considered to be consistent with findings from other empirical and modeling work, which similarly demonstrated that humans prefer to form representations based on a subset of the available dimensions (Lee, 2001; Shepard & Arabie, 1979; Sloutsky, 2003; Tversky, 1977; Ullman et al., 2002). Galdo et al. (2021), therefore, concluded that humans demonstrate a bias toward parsimonious solutions during learning, but nevertheless maintain some ability to flexibly allocate attention in order to improve performance.

We designed the within-trial module of AARM with these results in mind, given that the between-trial module is insufficient for explaining how humans decide when to terminate the information sampling process and commit to a choice during individual trials. The within-trial module predicts self-termination through a combination of stochastic feature imputation and bounded evidence accumulation. In the decision-making literature, accumulation-to-bound models specify mechanisms through which an observer samples information from a stimulus through time, and a response is made when evidence in favor of a particular choice exceeds a prespecified threshold. Unlike standard implementations (Brown & Heathcote, 2008; Ratcliff, 1978; Usher & McClelland, 2001) or extensions to multiattribute choice (Busemeyer & Townsend, 1993; Krajbich et al., 2010; Trueblood et al., 2014), however, AARM makes no assumption that moment-to-moment samples of information are independent, but rather are integrated with information from other dimensions to activate memories of exemplars and contribute evidence toward a category response. To determine which sources of information to sample, AARM first forms expectations about which features might occur in each dimension (based on past exemplars), and randomly draws potential feature values into a working representation of the stimulus. The observer then orients to dimensions that provide additional evidence in favor of the leading category option at each timestep and updates the working representation as features are encoded. This "confirmatory search" behavior naturally arises from the within-trial module's gradient-based mechanisms for updating attention, as will be discussed in the Attention as an Optimization Problem section below. For now, it is sufficient to establish that AARM continuously reorients attention to encode stimulus features into its working representation, and self-terminates when it samples enough information to surpass a decision threshold.

Although SEA's calculations are driven by predicted utility rather than a theoretical measure of attention, it is worth noting that SEA similarly does not impose a sum-to-constant constraint on its estimates of utility. The model instead implements parsimonious resource expenditure by (a) comparing the predicted utility of sampling a dimension to an expected cost and (b) limiting the depth of forward search when predicting utility (i.e., what Braunlich and Love (2021) refer to as a "mypoic"

rather than full preposterior analysis). Through ongoing utility calculations, SEA predicts self-termination when the potential gain of sampling any dimension no longer exceeds the potential cost of time and energy. Although this strategy is relatively efficient for low-dimensional stimuli, preposterior analysis requires the observer to determine the likelihood and category association of every possible combination of feature values across dimensions. This quickly incurs high-computational cost as more dimensions are added, even when using the myopic strategy of only making predictions one step into the future. AARM's approach, by contrast, incorporates human-like biases in the interest of approximating observed behavior. Its approach is therefore readily extendable to tasks involving higher dimensional stimuli, given that expected feature values are spontaneously retrieved from memory rather than being exhaustively considered.

In this way, AARM is similar to extensions to GCM that allow for sequential acquisition and retrieval of information. In the extended generalized context model (EGCM for response times [EGCM-RT] Lamberts, 2000), stimulus dimensions are sampled sequentially to facilitate gradual formation of a category representation through time. Similarly, the exemplar-based random-walk model (EBRW; R. Nosofsky & Palmeri, 1997) samples exemplars from memory and makes a decision when evidence surpasses a threshold. Unlike AARM, however, neither EGCM-RT nor EBRW has mechanisms for prioritizing dimensions according to task relevance, strategically reorienting to additional dimensions within-trial, or self-terminating the sampling process. Instead, both models sample and encode all available stimulus information before making a choice.

Attention as an Optimization Problem

Although GCM made a major theoretical contribution by relating attention to learning, an open question remained as to how attention should change as learning occurs. After a few early attempts to solve this problem (Estes, 1986; Gluck & Bower, 1988), perhaps the most complete theoretical description was provided by ALCOVE (Kruschke, 1992). ALCOVE combines exemplar-like representations used by GCM with an adaptive reinforcement policy engineered by a connectionist architecture. The model consists of three layers, connected by intervening sets of weights: an input layer contains the stimulus features, a hidden layer contains a set of exemplars, and an output layer contains the model's representation of a response probability. The set of weights that connect the latter two layers is referred to as "attention," given that they fulfill a similar purpose to the attention weights in GCM. As in the typical connectionist approach, back propagation is used to alter both sets of weights after each new experience by minimizing a loss function that compares the response probability output from the model to a vector representing the true category label (e.g., provided by feedback). Over time, adjustments to the attention weights minimize the total number of categorization errors. This updating process can be thought of as a first-order optimization process, solved by gradient descent. The intuition of the problem solved by ALCOVE is that attention weights should move to a location in the abstract, multidimensional attention space that minimizes the squared loss function over time. Later, a similar procedure was assumed by SUSTAIN (Love et al., 2004).

Although ALCOVE has many similarities to GCM, a major departure is that it does not allow for explicit storage of new episodic events as they are experienced. Instead, ALCOVE presupposes that a set of basis exemplars are specified prior to learning, and the connection weights between experiences and exemplars are adjusted through time. As clarified by Turner (2019), most learning models take one of two forms: an "instance" representation or a "strength" representation. The former consists of models that assume that each new experience is captured in episodic memory, creating an "instance" of the event (Estes, 1994; Logan, 1988, 2002; Medin Schaffer, 1978; R. Nosofsky, 1986). The latter consists of models that simply adjust a set of weights according to a rule, leaving no permanent storage of those events for future retrieval (D. Cohen et al., 1990; Rumelhart McClelland, 1988). Given this distinction, ALCOVE is a strength-based model because it learns by modifying its weight structures over time.

When making efforts to distinguish between these two classes of theories, one pervasive problem is the confound between attention and representation. Specifically, encoded information affects the representation of the feature-to-category map, and this representation can subsequently drive the deployment of selective attention. In this way, an introspective learner may wonder during a task "Am I attending this dimension because I have learned that it is relevant, or is this dimension only relevant because I have attended to it before?" Assuming a prototypical structure, strength-based models incur major theoretical limitations due to their lack of an explicit encoding structure for experienced events.

Recent research has begun to elucidate the interactions between the information that is stored, and the search for subsequent information. For example, Rich and Gureckis (2018) have shown that when only a subset of information is attended, subjects can fall into "learning traps" by inappropriately generalizing information to unattended dimensions. In other work, Turner et al. (2021) have shown that selective attention can cause subjects to falsely believe that one dimension is more relevant than it actually is, which can curtail a learner's willingness to explore new dimensions of information. These results suggest that increasingly selective deployment of attention across trials could be explained by the individual-specific history of encoded features and their learned relevance. The notion that attention orients based on an individual's "selection history" has become a popular way of thinking about how attention should be deployed in response to one's knowledge and one's goals (Awh et al., 2006). If we apply such logic in the context of category learning, there clearly becomes a need to specify which experiences enter into an observer's representation when determining how attention should orient.

To this end, AARM's within-trial module enables confirmatory information search, which is well documented in human learning (Lefebvre et al., 2022; Nickerson, 1998; Talluri et al., 2018). Using similar mechanisms to ALCOVE, AARM's between-trial module updates attention on each trial with respect to the correct category label, as provided by feedback. To extend the same mechanisms to account for sampling and decision dynamics, within-trial attention is first initialized with weights inherited from the previous trial. Attention is then updated at each timestep with respect to the category label that currently has the most evidence, given that the true category label is not known until after within-trial processes terminate in a response. As such, the observer reorients to dimensions that are expected to provide additional evidence for the

category that is believed to be correct, given the current state of knowledge about the stimulus.

By contrast, SEA specifies unbiased information search via preposterior analysis. The observer samples dimensions that are expected to serve the overall goal (i.e., increase the probability of making a correct category response) in excess of the potential cost. Broader sampling beyond the relevant dimensions is made possible in the model by adjustments to an exploration parameter. The distinction between confirmatory and unbiased information search exemplifies the differences between AARM and SEA and their respective purposes. For example, developmental work has shown that while adults tend to categorize new items according to a perfectly reliable dimension, children often make decisions in consideration of multiple dimensions with less regard for overall reliability (Blanco & Sloutsky, 2019; Deng & Sloutsky, 2015). While AARM can be used to identify which cognitive mechanisms potentially account for observed group-level differences in behavior, SEA can be used to assess the efficiency of the two strategies relative to rational predictions for behavior. Although confirmatory search in AARM provides a natural extension to between-trial mechanisms related error minimization to account for the unsupervised aspects of within-trial dynamics, this notable departure from optimal sampling may have limitations beyond the context of category learning. These potential limitations and directions for future investigation on its relevance to human behavior are addressed in the General Discussion section.

ALCOVE, SUSTAIN, AARM, and SEA all fundamentally specify learning as an optimization problem with respect to the observer's goals, but use different mechanisms to solve it. Given that the models make very clear predictions for how dimensions of information are attended over time in order to predict learning (Braunlich & Love, 2021; Galdo et al., 2021; Kruschke, 1992; Love et al., 2004; Mack et al., 2013, 2016), constraining and adjudicating between their respective theoretical assumptions potentially requires insights beyond what behavioral data alone can provide.

The Necessity of Eye-Tracking Data

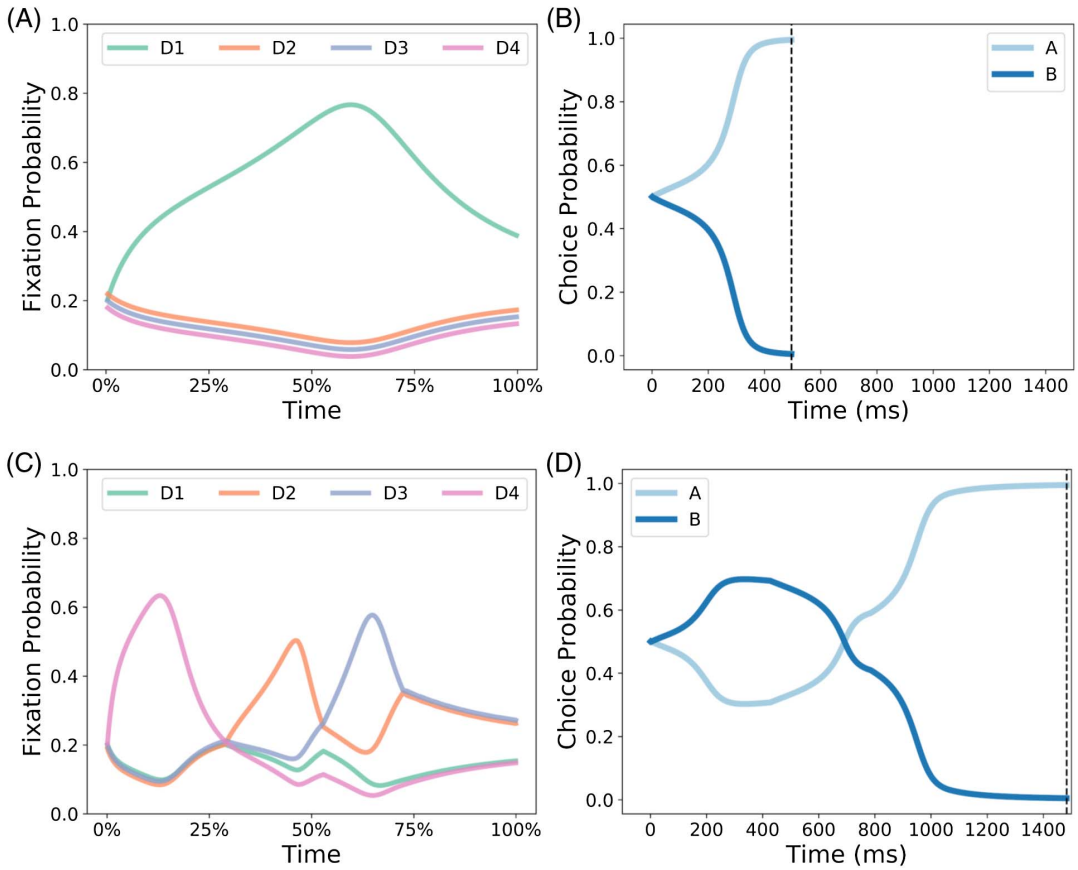
A central theme of this article is to use measures of gaze fixation as a guide for developing a model of category learning that considers both between- and within-trial dynamics. We are certainly not the first to use eye-tracking data to shed light on theories of category learning (see Lai et al., 2013, for review). To investigate the connection between latent and observable correlates of attention, Rehder and Hoffman (2005a) collected eye-tracking data while participants completed category learning tasks with different levels of complexity (Shepard et al., 1961). The authors demonstrated that eye-tracking data can distinguish among alternative model-based assumptions about how attention is allocated at the beginning of the task as opposed to the end after learning has occurred. ALCOVE (Kruschke, 1992), for example, predicts that observers initially distribute attention evenly across all dimensions before identifying which dimensions are most relevant. An alternative theory outlined by the rule-plus-exception model (RULEX; R. Nosofsky et al., 1994) assumes that observers implicitly form and test hypotheses during learning, and therefore predicts that observers would initially attend to a single dimension until its relevance could be sufficiently ascertained. It is important to note that these divergent assumptions could not have been examined with a measure as coarse as trial-level accuracy. One reason is that the distinction between pre- and postlearning was essential to the question of interest. Given that only the first trial contains information about attention in the absence of learning, using accuracy as the outcome measure would require conclusions to be heavily based on what is effectively a single data point. A second reason is that observers could use either an ALCOVE-like strategy of distributing attention evenly across dimensions, or a RULEX-like strategy of fixating on one dimension at random, and the predicted accuracy would be approximately equivalent on average. With eye-tracking data, however, Rehder and Hoffman identified fixation probabilities that were consistent with ALCOVE rather than RULEX: participants fixated to all dimensions with approximately equal probability at the beginning of the task and attended only to the most relevant dimensions toward the end.

While the results of Rehder and Hoffman (2005a) relied on trial-level fixation probabilities, additional evidence suggests that gaze fixation data can be used as a continuous measure of within-trial attention as well (Blair et al., 2009; Chen et al., 2013; Krajbich et al., 2010; Krajbich & Rangel, 2011; Rehder & Hoffman, 2005a; S. Smith & Krajbich, 2019a, 2019b; Thomas et al., 2019). In work by Blair et al. (2009), gaze fixation data were recorded while participants completed a category learning task with hierarchically organized stimulus dimensions (see Figure 1). As described in the introductory section of the current article, the feature value in one superordinate dimension indicated which of two subordinate dimensions would be relevant for determining the category label for each stimulus. If one were to fit a model like ALCOVE, SUSTAIN, or AARM's between-trial module to data from an experiment like this (see Palmeri, 1999, for an application of ALCOVE), we should expect the superordinate dimension to be preferentially weighted because it is relevant across all trials. The two subordinate dimensions would be weighted equally because they are each relevant to 50% of trials (Figure 1B). If one were to predict proportions of gaze fixations directly from these attention weights, one might expect a high probability of fixating to the superordinate dimension, and lower, but equal, probabilities of fixating to the two subordinate dimensions. In reality, Blair et al. (2009) noted distinct stimulus effects on the trajectory of within-trial fixations, such that participants conditionally fixated to only one subordinate dimension per trial after observing the feature identity of the superordinate dimension (see Chen et al., 2013; McColeman et al., 2014; Meier Blair, 2013, for replication). The results suggest that in addition to using learned information about dimension relevance to sample information, humans prioritize dimensions dynamically within a trial in response to the stimulus itself. Although Braunlich and Love (2021) demonstrated that SEA could predict a reduction in the number of dimensions sampled across learning instances, the computationally parsimonious "myopic" variant of SEA does not predict the ordering effects observed by Blair et al. (2009). Given that SEA considers all dimensions to have equal utility on average, it does not produce preferential orienting behaviors that are consistent with the hierarchical structure of the task. As we will show in Case Study 2, however, AARM's within-trial module can produce stimulus-dependent information prioritization effects through its combination of attention-mediated orientation and confirmatory information search.

In light of empirical and theoretical work indicating that the hierarchical organization of information is ubiquitous in human learning (e.g., Barto & Mahadevan, 2003; Botvinick, 2012; Botvinick et al., 2009), we suggest that the within-trial attention effects that emerge from hierarchical category structures can potentially make a more general statement about how humans sample information from naturalistic environments. For example, contextual features of the environment may serve as a set of superordinate dimensions for deciding which sources of information to attend when making judgments about new examples of recognizable objects that people encounter in everyday life. We, therefore, place particular emphasis on hierarchical category structures in the demonstrations of AARM's predictions in sections to follow.

As a final example to motivate the use of eye-tracking data in developing our theory of category learning, it is relevant to note that multiple modes of information sampling could yield inseparable patterns of behavior under certain conditions (Figure 3). Consider two hypothetical learners who are assigning four-dimensional stimuli to Categories A and B. One dimension (D1) is perfectly predictive of category membership, such that an observer could achieve 100% accuracy by learning the appropriate D1 feature-to-category mapping (e.g., when D1 = 1, respond Category "A," and when D1 = 2, respond Category "B"). The three other dimensions (D2, D3, and D4) are each 75% predictive of category membership. Learner 1 is very efficient and identified the most reliable dimension and exclusively sampled information from D1 after gaining

Figure 3
Information Sampling and Decision Dynamics



Note. AARM = adaptive attention representation model. Hypothetical fixation paths were generated by AARM's within-trial module, such that one of four spatially segregated dimensions was fixated at each timestep up to a response. Left panels show the probabilities of fixating to each dimension (*y*-axis), plotted as a function of the percentage of time within trial between stimulus onset and response (*x*-axis). Right panels show the decision evidence for each of two possible category choices as a result of the information sampling behavior (i.e., fixation paths) in corresponding left panels. Choice probability (*y*-axis) is plotted as a function of absolute time in milliseconds (*x*-axis). Dotted lines indicate when self-termination (i.e., a response) occurred. Each row shows the timecourses of fixations and decision evidence for: (A) a hypothetical subject who learned to attend to the deterministic (100% predictive of category; D1) dimension; (B) a hypothetical subject who received conflicting evidence across three probabilistic dimensions (D2, D3, and D4). Although each simulation reflects different information sampling behaviors, Category A was selected in both examples. See the online article for the color version of this figure.

experience with the task. In Figure 3A–B, we show an example in which a learner fixated to D1, and concurrently accumulated considerable evidence to support a Category "A" decision.

By contrast to Learner 1, Learner 2 happened not to notice that D1 was the most reliable dimension. Instead, they found that by attending to some combination of D2, D3, and D4, they could achieve very high accuracy that in fact rivaled that of Learner 1. Figure 3C–D shows an example in which a learner prioritized D4, which provided some initial evidence for a "B" response. Sampling information from D2 subsequently contradicted the information in D4 and created uncertainty in the choice. To resolve this conflict, the learner sampled information from D3, which provided sufficient information for making an "A" response. Given that these two divergent learning profiles could yield identical accuracy, responses alone may say very little about whether or not a model is accurately capturing which dimensions are attended. While different modes of information sampling could be dissociated by clever task design (e.g., Blanco & Sloutsky, 2019; Deng & Sloutsky, 2015), measures of attention such as those provided by eye-tracking data provide strong constraints on how attention is deployed over time within a trial. In particular, a viable model of category learning that uses latent attention to predict behavioral changes should be able to account for multiple modes of observable attention allocation as well. We will further explore the impact of different sampling paths on response probability in Case Study 1.

The examples highlighted in this section seek to clarify that there are at least two problems in using behavioral data as the lone metric for validating the assumptions of attentional deployment. First, when stimuli are multidimensional, it is possible for many patterns of attention allocation to produce identical responses. Rehder and Hoffman (2005a) showed that eye-tracking data could be used to support a broad distribution of attention early in the learning period, as opposed to a systematic testing of one dimension at a time. Relatedly, Figure 3 illustrated how different sequences of fixation patterns within a trial could ultimately produce the same category choice. Second, dimension relevance may be highly contextualized within a trial, based on the properties of the stimulus itself. A particularly striking example comes from hierarchical category learning experiments in which participants fixate to dimensions in a stimulus-dependent manner before feedback is even observed (Figure 1, Blair et al., 2009).

Despite overwhelming evidence that eye-tracking data provide a rich source of information about the timecourse of selective attention during individual decisions, few efforts have been made to extend the logic of categorization models to account for within-trial dynamics (but see Braunlich & Love, 2021). Taking these findings together, we assert that gaze fixations serve as a viable, necessary means for evaluating category learning models in terms of predicted attention allocation. By using eye-tracking data in the current work, we are equipped to examine the theoretical mechanisms put forth by AARM using a new standard of specificity, to which other models of category learning have been infrequently subjected.

Summary and Outline

The introductory sections have supported the notion that attention is a critical component for learning problems: It accelerates learning by increasing the influence of relevant dimensions on decision processes, and thereby limits the time-consuming search for information when making decisions. Our conceptualization of how attention should be deployed follows those of Kruschke (1992), Love et al. (2004), and Galdo et al. (2021) by treating attention as an optimization problem. At face value, the problem of optimizing attention should be similar at the between- and within-trial level, but there are important differences that make for an interesting challenge. Across trials, the learning problem is well-defined: One needs only to specify how attention should be modified in response to feedback (i.e., supervised learning). However, within a trial, the problem is made more complex because the learner does not know the true category label until after they make a response, but must nevertheless decide which dimensions to sample (i.e., unsupervised learning).

The gold standard in solving the problem of unsupervised learning is some type of forward computing, whereby relevance is determined by considering all possible values of a dimension and then aggregating the results to form an expected utility of each (Nelson & Cottrell, 2007; Yang & Lengyel, 2016). SEA is perhaps the most striking display of this approach, in which the utility of sampling is computed for all dimensions prior to making a decision to act (e.g., sample a dimension or make a response). Although this approach has considerable promise, one potential weakness is that it assumes an incredible amount of computation at each moment in time to assess the potential utility of every sampling outcome. It is possible that humans do indeed make these computations, but it certainly is not an economical approach if suitable heuristic alternatives were available, particularly in consideration of high-dimensional stimuli.

By contrast, AARM focuses on the representation of the current stimulus information rather than on the utility of would-be collected information. The rationale behind this strategy is that subjects maintain a sense of the distribution of features that occur within each dimension, and they use this distribution to form expectations about the current stimulus. By dynamically updating the expectations, a "working" representation can subserve attention and the search for subsequent information. To solve the unsupervised aspect of this problem, we critically assume that information accumulates in a confirmatory manner until category evidence surpasses a decision threshold. This assumption appears to be vital to our approach, as it naturally extends the between-trial module of AARM (Galdo et al., 2021) to account for within-trial dynamics. To articulate our proposed framework, we consider how latent attention is updated between trials to facilitate learning, and within trials to facilitate individual categorization decisions. Mathematical details and justification will be provided in the sections to follow, but our theoretical framework can be summarized by the following core components:

- Both within- and between-trial dynamics are described by a common set of mechanisms. Interactions among attention, representations, and decisions extend across timescales to account for how humans acquire information about individual stimuli and learn how to distinguish between categories.

- Over the course of learning, humans form simplified stimulus representations composed of the dimensions that are most relevant to the current task. Within-trial dynamics of information sampling and decision-making describe how these simplified representations are formed, such that only a subset of information needs to be attended before a categorization response is made.
- Attention is optimized with respect to the current goal. Gradient-based mechanisms typically require the observation of feedback to update the attention weighting structure between trials. If we extend the same logic to the within-trial level, we must define how the observer orients attention before the correct category label is known. We therefore describe how representations gradually evolve within trial according to experience-based predictions and confirmatory information search.
- Hierarchical category structures are ideal for studying within-trial dynamics, due to an implicit temporal ordering of relevant information. In addition to giving rise to gaze prioritization effects in an experimental setting, hierarchical structures are ubiquitous in nature. We suggest that in real-world scenarios, learners use environmental context as a superordinate cue for processing the dimensions of new stimuli.

To explicate these theoretical components, the remainder of this article is organized as follows. First, we discuss the mathematical details of AARM in terms of two separable but interacting modules. We begin with a description of how AARM is applied to between-trial learning, and then describe how expectations about features can be managed dynamically to create a working representation of the stimulus probe. We then describe how attention orients to confirm the existing beliefs about a stimulus to complete our description of within-trial dynamics. Second, we examine AARM's ability to capture important empirical effects by simulating its behavior in four case studies. The case studies examine how attention is deployed in several unique situations: (a) when expectations are violated, (b) when relevance is contextualized within a stimulus (e.g., as in the hierarchical category learning task), (c) when multiple stimulus dimensions occupy the same location in space, and (d) when learning occurs incidentally (e.g., dimensions are not relevant to the learning process but become relevant when tested). We close the article with a discussion about future directions and alternative mechanisms.

Model Specification

To present the details of AARM, we separate our description into distinct between- and within-trial updating processes as shown in Figure 2. First, as described by Galdo et al. (2021), the model updates the category representation in response to feedback on each successive trial (i.e., a between-trial update). This type of update entails (a) storing a new episodic trace containing the stimulus information on the current trial, (b) storing information about the category label (e.g., from feedback), and (c) updating the quantity of attention that is allocated to each dimension.

Second, the model maintains a representation of the current stimulus probe, which it updates through time as it encodes new information about each feature (i.e., a within-trial update). This type of update entails (a) an encoding process where attention is applied to a stimulus dimension in order to access the feature value contained therein, (b) an imputation process where the model uses the available information to form expectations about which feature values will occur in unattended dimensions, and (c) an attention rule that allows the model to reorient according to its updated knowledge and expectations about the current stimulus probe. We begin with a general overview of the between-trial module and expand this model structure to accommodate within-trial dynamics. For reference, a notation table and parameter definitions are provided in Appendix B.

Between-Trial Updating Rule

The relevant mechanisms of AARM's between-trial updating rule will be provided here, but we refer the reader to Galdo et al. (2021) for additional details. On each trial $\mathbf{i}$ of a categorization task, the observer is asked to assign a D-dimensional stimulus $\mathbf{e^{(i)}}$ to one of C categories. To do this, the observer is thought to retrieve memories of previously experienced exemplars $\mathbf{X^{(i)}} = [\mathbf{x^{(1)}} \ \ldots \ \mathbf{x^{(i)}} \ \ldots \ \mathbf{x^{(N_i)}}]^{\mathrm{T}}$ and their associated category labels $\mathbf{F}^{(i)} = [f^{(1)} \ \ldots \ f^{(i)} \ \ldots \ f^{(N_i)}]^{\mathrm{T}}$ (i.e., as supplied by corrective feedback). As in GCM, AARM assumes that memories of stored exemplars are "activated" in proportion to their similarity to the current stimulus. Similarity is computed by way of a factorizable exponential kernel (R. Nosofsky, 1986; Shepard, 1987), such that activation $a^{(n)}$ of the nth exemplar in response to probe $\mathbf{e^{(i)}}$ is as follows:

$$a^{(n)} = \exp\left(-\delta_B \sum_{j=1}^{D} \alpha_j^{(i)} |e_j^{(i)} - x_j^{(n)}|\right) m^{(n)}. \tag{1}$$

Here, δ_B is the specificity of the between-trial similarity kernel function, $\mathbf{M^{(n)}} = [\mathbf{m^{(1)}} \ \ldots \ \mathbf{m}^{(\mathrm{N_i})}]^{\mathbf{T}}$ contains the memory strength associated with each exemplar, and $\boldsymbol{\alpha}^{(i)} = [\alpha_1^{(i)} \ \ldots \ \alpha_D^{(i)}]$ quantifies the attention allocated to each of the D stimulus dimensions. Although Galdo et al. (2021) assumed M was determined by a weighting function that incorporates primacy and recency biases (Pooley et al., 2011), we assumed all exemplars had equivalent memory strength to provide constraint in our simulation case studies. The probability of choosing Category c is the summed similarity of the exemplars associated with that category, normalized by the total across all exemplars (i.e., a weighted average). Specifically, the choice probability $V_c^{(i)}$ associated with Category c is as follows:

$$V_c^{(i)} = \frac{\sum_{n=1}^{N} a^{(n)} \mathbb{I}\left(f^{(n)} = c\right)}{\sum_{n=1}^{N} a^{(n)}}, \tag{2}$$

where $\mathbb{I}(f^{(n)} = c)$ is an indicator function returning a one if the statement is true and a zero otherwise.

After a response is made and feedback is observed, two actions occur. First, the features of stimulus $\mathbf{e^{(i)}}$ are stored in exemplar matrix $\mathbf{X}$ as a memory trace, and the true category label is stored in feedback matrix $\mathbf{F}$. Second, attention $\boldsymbol{\alpha}^{(i)}$ is updated in the direction

of an error gradient, similarly to the adaptive attention models described in the Attention as an Optimization Problem section (i.e., ALCOVE and SUSTAIN):

$$\alpha^{(i+1)} = \alpha^{(i)} - \gamma_B \nabla_\alpha \text{loss}\,(\alpha^{(i)}). \tag{3}$$

Here, γ_B is a positive constant describing a between-trial learning rate, and ∇_α is a shorthand denoting a "gradient operator" for computing the set of partial derivatives of a loss function $f(\alpha)$ with respect to each element of the vector α:

$$\nabla_\alpha f(\alpha) := \left[\frac{\partial}{\partial\alpha_1} f(\alpha) \quad \frac{\partial}{\partial\alpha_2} f(\alpha) \quad \ldots \quad \frac{\partial}{\partial\alpha_D} f(\alpha)\right].$$

To define the loss function, ALCOVE and SUSTAIN use the so-called "humble teacher" rule, which allows for variability in category activation between exemplars. Specifically, more category-typical exemplars elicit greater activation than those that are more peripheral (Kruschke, 1992). For the purposes of our previous work on simplicity biases in human learning (Galdo et al., 2021), we instead selected a cross-entropy loss function because it allows for faster training and more reliable extension to multiclass problems than squared-loss alternatives (Demirkaya et al., 2020). Given successful fits to behavioral and eye-tracking data with our previous specification for between-trial attention updating, we apply cross-entropy loss in the current work as well.

When using a soft-max rule (such as Luce-choice), the cross-entropy loss function is simply the negative log-likelihood of correct classification (Goodfellow, 2016):

$$\nabla_\alpha \text{loss}\,(\alpha^{(i)}) = -\nabla_\alpha \log\left(V^{(i)}_{f^{(i)}}\right),$$

where $V^{(i)}_{f^{(i)}}$ is the choice probability associated with the feedback given on the ith Trial (i.e., the correct response). Hence, to derive the gradient, we need only take the partial derivative of Equation 2 with respect to $\boldsymbol{\alpha}^{(i)}$ along each of the D dimensions. We provide this derivation in Appendix A. Our update equation for the attention vector $\boldsymbol{\alpha}^{(i)}$ after observing feedback $f^{(i)}$ therefore becomes:

$$\alpha^{(i+1)} = \alpha^{(i)} + \gamma_B \nabla_\alpha \log\left(V^{(i)}_{f^{(i)}}\right). \tag{4}$$

We stress that this updating procedure departs from strength-based connectionist architectures in which back propagation solutions provide the rule to update both the attention vector $\boldsymbol{\alpha}^{(i)}$ and hidden layer weights. As shown by Galdo et al. (2021), simply defining how attention should be updated across time within an instance representation is sufficient to predict human categorization behaviors.

Within-Trial Updating Rule

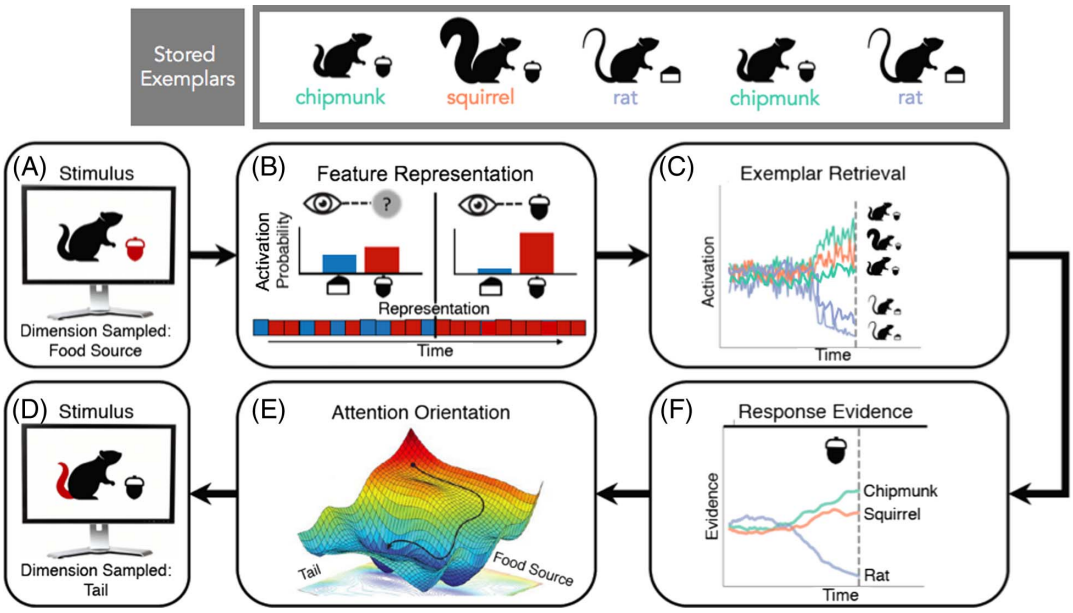
Figure 2B illustrates the important components of the within-trial updating process, as well as its "default" temporal order (i.e., nodes with numbers). Upon stimulus presentation, a set of initial attention weights inherited from the previous trial of the between-trial module dictates which information will be sampled. In this process, the eyes orient to the location of the prioritized dimension (Node 1 in Figure 2B). Once the eyes have fixated upon the intended dimension, a feature encoding process begins. Feature information is then passed to the representation (Node 2 in Figure 2B), and similarity-based activation of the stored exemplars is used to calculate evidence for each category response (Node 3 in Figure 2B). The model reorients attention in a confirmatory manner, according to which dimension is most likely to provide further evidence that would support whichever choice currently has the largest amount of supporting evidence. This dynamic process self-terminates and makes a response when a sufficient amount of evidence has accumulated for an option.

For ease of exposition, we organize this section into the following four stages of processing: stimulus encoding, exemplar activation, evidence for category response, and attention orientation. Figure 4 provides an illustrative example of how each of these components contributes to within-trial dynamics, and we will use this figure as a working example to facilitate descriptions of each component.

Stimulus Encoding

Memory theories often describe the psychological representations of stored items or events as memory "traces," which are organized into discrete features of perceptual, contextual, and conceptual information. While the contents of a memory trace cannot be directly observed, recall and recognition paradigms provide insight into which features are encoded under various conditions. For example, if a lure item is falsely recognized among previously studied targets at test, it indicates overlap between the features of the lure and some subset of target memory traces (Deese, 1959; Roediger & McDermott, 1995). Additional work has shown that the distribution of features across stored traces and the extent to which they can be associated with one another influence which information will be encoded and subsequently retrieved (Dosher, 1984; Dosher & Rosedale, 1991; Greene & Tussing, 2001). With these insights in mind, a recent dynamic model of encoding and retrieval (Cox & Criss, 2020; Cox & Shiffrin, 2017) described trace formation as a time-varying process. In this account, the iterative encoding of individual probe features selectively activates memory traces on the basis of similarity, and drives an evolving familiarity signal toward a recognition threshold. Similarly, AARM's within-trial module was designed to build up an informative representation of the probe throughout the trial, using retrieval of previous exemplars to drive an evidence accumulation signal for making a category response. Our specification of encoding in AARM's within-trial module builds upon mechanisms of prediction and pattern completion observed in the hippocampus, in which previously observed item representations are reinstated during encoding in order to fill-in missing information or properly orthogonalize overlapping cues (see Bowman Zeithamova, 2020; Hunsaker Kesner, 2013, for review). Prior to encoding any information about a new stimulus, we assume that a working representation is populated by experience-based expectations of feature values. Expectations are gradually replaced with the true features of the stimulus as they are attended and concurrently encoded over the course of the trial.

Figure 4
Illustration of Within-Trial Dynamics



Note. (A) An example stimulus is presented on the screen, and a stimulus dimension is sampled for processing (e.g., prioritized from the between-trial module). (B) The observer generates a working representation of the stimulus and predicts what features might occur in each dimension. As a feature is attended, predictions are replaced with true feature values. (C) Previously stored exemplars are activated in proportion to their similarity to the probe. (F) The category labels associated with retrieved exemplars accrue noisy response evidence. (E) Attention updates to discriminate among the currently most active category options. (D) Gaze fixations are determined from the attention process, resulting in reorientation to new dimensions as needed to sample more category-relevant information. See the online article for the color version of this figure.

To incorporate the logic of pattern completion into AARM's encoding mechanism, we follow a procedure outlined by Estes (1994). Borrowing his example, suppose an observer experiences the following three-dimensional stimuli: $\mathbf{e^{(1)}} = [\mathbf{1}, \mathbf{1}, \mathbf{1}]$, $\mathbf{e^{(2)}} = [\mathbf{1}, \mathbf{2}, \mathbf{1}]$, and $\mathbf{e^{(3)}} = [\mathbf{1}, \mathbf{2}, \mathbf{2}]$. Further suppose that on Trial 4, the observer is presented with a partial stimulus $\mathbf{e^{(4)}} = [\mathbf{1}, \mathbf{1}, ?]$ and is asked to guess which feature value will occur in Dimension 3. We assume that an observer will predict the feature value based on memories of previous items and the current state of knowledge about the stimulus. To make and evaluate predictions, one can impute each feature value that was previously observed in Dimension 3 (i.e., [1, 2]) into the partial stimulus, and evaluate which feature value is more likely to represent the missing information.

Starting arbitrarily with a candidate feature value of 1 as shown in Table 1 below, we compare the imputed stimulus (i.e., 111) to all stored exemplars and determine whether the feature values match or mismatch in each dimension. In the *Comparison* column of Table 1, "matches" and "mismatches" are indicated by values of 1 and y respectively, where y represents a baseline level of perceptual discriminability. We then compute the product across comparison values to determine the *Similarity* column (Medin & Schaffer, 1978). Finally, we compute the sum similarity across all stored exemplars to determine the activation of imputed stimulus 111.

Similarly, we can calculate the activation when "2" is the missing value using the same strategy (Table 2).

The probability of selecting a value is simply the activation of its respective imputed stimulus, normalized by the total activation across all candidates. In our example, the probability that the stimulus $\mathbf{e^{(4)}}$ has a feature value of 1 in Dimension 3 (i.e., $e_3^{(4)}$) is as follows:

$$P\left(e_3^{(4)} = 1\right) = \frac{1 + y + y^2}{1 + 3y + 2y^2}.$$

As long as y is small enough to indicate sufficient perceptual discriminability among candidate feature values, $P(e_3^{(4)} = 1)$ will approach 1. In other words, when asked to complete the partial stimulus $\mathbf{e^{(4)}} = [\mathbf{1}, \mathbf{1}, ?]$, the observer is most likely to respond "1."

In extending the intuition of Estes's example for the purposes of dynamic encoding, it is necessary to distinguish between the "true" identity of the stimulus and a "working" representation that changes through time. As in our description of the between-trial module, we use the notation $\mathbf{e^{(i)}}$ to denote the true identity of the probe on the ith trial. We denote the working representation of the stimulus probe at Timestep t of Trial i as $\mathbf{e}^*(t) = [e_1^*, \ldots, e_D^*]$ and omit the "i" trial notation for convenience. We next require a general expression for the probability that a candidate feature value will occur in a particular dimension. We define the set of H unique feature values that were previously observed in Dimension j as $\mathcal{R}_j = [r_1, r_2 \ldots r_H]$. We then use the following equation:

$$z_j^{(n)}(t, r_h) = \exp\left(-\delta_W \alpha_j^*(t) |r_h - x_j^{(n)}|\right) m_j^{*(n)}, \tag{5}$$

Table 1
Projected Similarity (1)

Imputed stimulus	Stored exemplars	Comparison	Similarity
111	111	111	1
	121	1*y*1	y
	122	1*yy*	y^2
Sum:			$1 + y + y^2$

to calculate activation $z_j^{(n)}(t, r_h)$ of stored feature value $x_j^{(n)}$ in response to an imputed feature value r_h, drawn from $\mathcal{R}_j$. This equation is a more general form of the exemplar activation calculation provided in Equation 1. Here, however, δ_W is the within-trial specificity of the similarity kernel function, $\boldsymbol{M}^{*(n)} = [m^{*(1)} \ldots m^{*(N)}]^{\mathrm{T}}$ indicates the encoding status of exemplar features, and $\alpha_j^*(t)$ indicates the attention weight at moment t. We will describe how attention changes through time in the Attention Orientation section, but for now it is sufficient to acknowledge that attention is updated throughout the trial and will affect how stimuli are encoded.

Using the relation $\exp\left(\sum_i x_i\right) = \prod_i \exp(x_i)$, the probability that the "true" feature value $e_j^{(i)}$ is equal to r_h is as follows:

$$P(t)\left(e_j^{(i)} = r_h\right) = \frac{\sum_{n=1}^{N}\left[z_j^{(n)}(t, r_h)\prod_{k\neq j} z_k^{(n)}(t, e_k^*(t))\right]}{\sum_{s\in\mathcal{R}_j}\sum_{n=1}^{N}\left[z_j^{(n)}(t, s)\prod_{k\neq j} z_k^{(n)}(t, e_k^*(t))\right]}. \quad (6)$$

Note that Equation 6 takes the same form as the feature probability calculation from Estes's example. Here, the numerator is simply the activation associated with an imputed feature value r_h, and the denominator is the total activation associated with all values in $\mathcal{R}_j$. To specify a feature value in the working representation $e_j^*(t)$ at moment t, we randomly draw a value from the distribution defined by the probability mass function in Equation 6. Importantly, a new value of e_j^* is redrawn at each timestep within the trial, such that the working representation is nonstationary. This imputation process continues until sufficient attention has been applied to Dimension j for a true feature value to be encoded, at which point $e_j^{(i)}$ is predominantly represented in e. Although Equation 6 expresses the pattern matching probabilities for discrete feature values, it can easily be extended to continuous values by replacing the summation over the set $\mathcal{R}_j$ to be an integration over the space $\mathcal{R}_j$, in the same way that the similarity kernel (e.g., Equation 5) was generalized from Medin and Schaffer's (1978) context model to R. Nosofsky's (1986) GCM. We demonstrate an extension to a paradigm with continuously valued dimensions in Case Study 3.

Table 2
Projected Similarity (2)

Imputed stimulus	Stored exemplars	Comparison	Similarity
112	111	11*y*	y
	121	1*yy*	y^2
	122	1*y*1	y
Sum:			$2y + y^2$

Our specification of a prediction-based working representation is somewhat related to utility predictions in SEA. Both models assume the observer maintains an ongoing sense of what features might occur in each dimension, with an associated likelihood of occurrence that depends on the state of knowledge about the current stimulus. One critical distinction is how each model uses these insights to decide which sources of information to sample. While SEA requires a pairwise assessment of every possible combination of features in order to determine a single utility prediction for each dimension, AARM's working representation is more reflective of spontaneous, noisy retrieval of features that are unbounded by specific exemplar representations. As such, our approach has a similar intuition to Monte Carlo algorithms in which probability distributions are approximated through repeated sampling, some specifications of which can be recursively updated as more information is obtained (Doucet et al., 2001; Gilks et al., 1996). Random sampling approaches have been suggested to provide an advantage of cognitive plausibility over rational models on the grounds of computational parsimony (Sanborn et al., 2010). Relative to SEA, the feature imputation strategy in AARM is arguably more consistent with the capabilities of resource-limited humans because there is no requirement that every possible feature combination is assessed within the working representation. Expected or observed feature values are instead drawn from a distribution, and attention and decision components are updated accordingly.

Our specification is also similar to other extensions to GCM that were designed to characterize the timecourse of stimulus encoding during category learning tasks (Brockdorff & Lamberts, 2000; A. Cohen & Nosofsky, 2003; Lamberts, 2000). As mentioned previously, the EGCM-RT (Lamberts, 2000) incorporated a stochastic stimulus representation mechanism into GCM, which resulted in a similarity output that changes throughout the trial as probe dimensions are encoded. Unlike AARM, however, EGCM-RT does not specify a precise order in which dimensions should be encoded, only that encoding is sequential and that all feature values of the stimulus need to be encoded before a response is made. A variant of EBRW (R. Nosofsky & Palmeri, 1997) for perceptual encoding (EBRW-PE; Cohen & Nosofsky, 2003) contains similar stochastic dimension-sampling mechanisms, such that exemplars race toward a threshold at rates that are proportional to their total similarity to the probe. At each timestep within a trial, there is an increasing probability that a feature will be encoded and thus included in the continuous similarity calculation. As such, encoding a feature value within the stimulus representation is strictly probabilistic. By contrast, AARM offers a mechanism for encoding individual feature values that are driven by attention and are gated by gaze fixations.

Instead of populating the working representation with random draws from an expected distribution of feature values, an alternative approach would have been to define the working representation as an empty vector prior to encoding. The retrieving effectively from memory model (REM; Shiffrin & Steyvers, 1997), for example, assumes that observers begin with an empty trace vector of zeros. Over time, the zero elements of the trace are replaced with samples from a prespecified distribution (e.g., a Geometric distribution) with properties intended to reflect the details of the stimulus set. In the context of a model designed to capture within-trial dynamics, however, we found that the expectation-formation component of the working representation was essential for the model to reorient to additional dimensions after processing the first. In hierarchical

paradigms like the one illustrated in Figure 1 (Blair et al., 2009), various iterations of the model in which the stimulus representation was initialized with an uninformed (e.g., zero or average) basis vector provided no impetus for the model to reorient to one dimension over the other. As we will show in Case Study, our implementation achieves human-like reorientation to the stimulus-relevant subordinate dimension by updating its feature predictions after initial encoding, and fixating to a second dimension through confirmatory search.

Figure 4B illustrates how the encoding dynamics occur in AARM's within-trial module after initial orientation to a dimension (e.g., food source; Figure 4A) when an observer is categorizing images of animals. Before a new image is even presented, the observer has some expectation about what feature values each dimension could possibly take on, given their experience with previous stimuli. After the food source dimension is sufficiently attended and the observer encodes the "true" feature value (e.g., acorn), the working representation of the stimulus is updated to accommodate this information. As shown in Figure 4C and discussed below, this shift in the probe representation directly affects which stored exemplars are subsequently activated to facilitate the reorientation of attention.

Exemplar Activation

We assume that encoding (and by extension, attention) is the primary mechanism driving the activation of previously stored exemplars. This is in contrast to EBRW (R. Nosofsky & Palmeri, 1997) which assumes that the similarity of previously stored exemplars to the stimulus probe is what dictates how frequently each exemplar is retrieved. In AARM, attention is what guides the similarity computation itself, causing rapid nonlinear activation in both the activation of past exemplars and the evidence for a category response.

Exemplars are activated in a nearly identical way as described in the between-trial case (see Equation 1), with the one exception that activation is based on the working representation of the stimulus probe, $e^*(t)$, and not the true contents of the stimulus probe itself. In addition, activation is expressed as a function of time, given by the following equation:

$$a_n(t) = \exp\left(-\delta_W \sum_{j=1}^{D} \alpha_j^*(t)\left|e_j^*(t) - x_j^{(n)}\right|\right) m^{(n)}, \tag{7}$$

where we denote the attentional state at Time t as $\alpha_j^*(t)$.

As discussed in the previous section, the working stimulus representation in AARM's within-trial module is nonstationary and gradually comes to resemble the stimulus's true identity as features are encoded. As a consequence, the distribution of expected feature values in Equation 6 will change dynamically through time and affect which dimensions are prioritized, given the information available at Time t. Pertaining to the hierarchical paradigm shown in Figure 1 (Blair et al., 2009), Figure 5 shows how attention and exemplar activation mutually impact one another. At the beginning of the trial, memories for all exemplars are equally active (left panel, $t = 1$). Attention initially orients to the D1 dimension (right panel; x-axis) per weights inherited from the between-trial module. As the working representation is updated with D1 feature information, there is a concurrent retrieval bias for exemplars belonging to "A" categories (left panel, $t = 2$ and $t = 3$). When attention then updates again, the observer will reorient to D2 in an effort to distinguish between the categories associated with the most active exemplars (right panel; y-axis). When sufficient attention is applied to encode the feature value of D2, exemplars with similar features in both D1 and D2 are selectively activated (left panel, $t = 3$ and $t = 4$).

To account for potentially imprecise mappings between the visual properties of matching probe and exemplar features, we incorporated the notion of perceptual variability into the calculation for exemplar activation (Equation 7). As it stands, the distance calculation $|e_j^*(t) - x_j^{(n)}|$ within Equation 7 assumes the observer will perceive all matching feature values in a precisely identical way. This is unlikely to be the case for human subjects, whose visual perception depends on noisy neuronal firing and elements of bottom-up salience. For example, stimuli like Gabor patches (see Case Study 3) that allow for continuous dimensions of frequency and tilt angle are unlikely to be mapped to their true feature values such that all stimuli are precisely distinguishable. When attempting to generate human-like behavior in our case studies, we therefore added random noise drawn from a normal distribution with standard deviation σ to the distance calculation at each Time t. SEA provides an alternative method for accounting for imprecise feature perception and storage, in that memories are stored as clusters (Anderson, 1991a) and the likelihood of a stimulus belonging to each cluster is represented as a continuously updated distribution of belief. The probability that a given stimulus belongs to a particular category is determined by a weighted combination of cluster probabilities. Uncertainty is therefore inherent to the Bayesian belief-updating process in SEA, whereas AARM assumes precise mappings between stimulus and exemplar features unless otherwise specified (e.g., with noise).

In summary, exemplars that are more similar to the working representation of the probe will be more strongly activated in AARM and will have more relevance to the response choice. As a simplification of Figure 5, Figure 4C provides an illustration of how exemplar activation occurs. After the "acorn" feature is encoded as the food source of the probe stimulus, memories for animals that eat acorns are selectively activated, (e.g., chipmunks and squirrels) whereas memory traces for animals that eat cheese are deactivated (e.g., rats). In the next section, we will discuss how exemplar retrieval manifests in the accumulation of evidence in favor of an available category label.

Evidence for Category Response

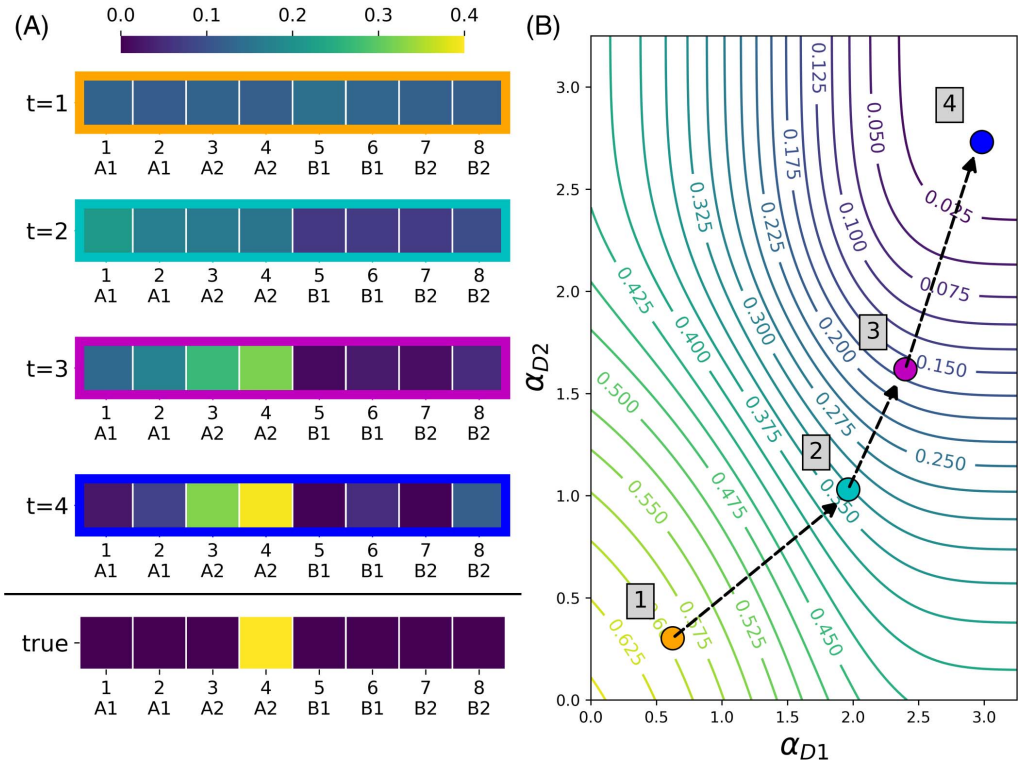
We assume within-trial choice probability is calculated in a way that mirrors the between-trial case (see Equation 2). Here, however, we reconceptualize "choice probability" as "decision evidence," and specify evidence at each time point using the following equation:

$$V_c^{(i)}(t) = \frac{\sum_{n=1}^{N} a_n(t)\,\mathbb{I}\left(f^{(n)} = c\right)}{\sum_{n=1}^{N} a_n(t)}, \tag{8}$$

where $a_n(t)$ is defined in Equation 7. The numerator represents the activation of the subset of exemplars associated with feedback c, and the denominator represents the total activation across all stored exemplars.

Figure 5

Illustration of Attention Gradient



Note. (A) Heatmaps show the activation of each unique exemplar in the task paradigm shown in Figure 1. *Y*-axis labels show trial numbers and the feedback associated with each exemplar. Activation at four different time points within an individual trial is shown, given a probe with a true category label of A2. (B) The plot shows the progression of within-trial attention weights assigned to dimensions D1 and D2, which are the relevant dimensions for determining the category membership of the given stimulus. As time progresses (as indicated by black arrows), the attention weights (*x*- and *y*-axis values) move in a direction to support a category response (contour values). See the online article for the color version of this figure.

EGCM-RT (Lamberts, 2000) uses a similar calculation to Equation 8 to approximate within-trial dynamics, and the output is interpreted as a probability of making each possible response, given an observed RT. In order to implement self-termination behavior into AARM's within-trial module, we instead assume that $V_c^{(i)}(t)$ for each Category *c* represents category evidence that accumulates up to a threshold. Alternative specifications would have been to use a race-like structure and calculate decision evidence as the sum of category-relevant exemplar activation without normalization (Brown & Heathcote, 2008; Usher & McClelland, 2001), or apply a log-ratio calculation similar to the sequential probability ratio test (SPRT; Wald & Wolfowitz, 1948). For our purposes of extending mechanisms for between-trial learning to account for within-trial sampling dynamics, however, it was important to use the same specification of the choice rule in order to ensure predictable behavior of the gradient-based attention update that will be discussed in the next section.

In specifying a decision rule, several options were available for consideration. We adopted a simple relative decision rule, such that the difference between the response with the largest evidence minus the response with the second-largest evidence must be greater than some value, ϕ. This specification is similar to extensions of the drift-diffusion model (DDM; Ratcliff, 1978) and the SPRT for modeling multialternative choice, where the decision terminates according to a thresholded distance between the two leading outcomes (McMillen & Holmes, 2006). Other approaches would have been to apply an absolute decision rule in which the threshold was applied to evidence for the leading option, or to apply a threshold to the distance between evidence for the leading option and the average evidence across alternatives (Niwa & Ditterich, 2008). Through testing, we found that our chosen specification provided the most stringent requirement for category-disambiguating evidence, such that multiple sources of task-relevant information were consistently sampled before decisions were made in a manner that was reflective of observed behavior.

Returning to Figure 4, Panel F shows how exemplar activation dynamically affects category evidence. In the example, the observer maps the most active exemplars to the "squirrel" and "chipmunk" categories, respectively, thus increasing the probability of making a "squirrel" or "chipmunk" response. The probability of responding "rat" concurrently decreases, given that the corresponding exemplars do not match the current stimulus with respect to the encoded

food source: acorn. As we will describe in the next section, category evidence at each timestep is used to continuously update attention and concurrent information sampling behaviors.

Attention Orientation

On a between-trial basis, the attention vector $\boldsymbol{\alpha}^{(i)}$ reflects the quantity of attention deployed to each stimulus dimension on the *i*th trial. We assume that attention in the within-trial module initially orients according to learned experiences, which are synthesized by these between-trial attention weights. For example, if a subject learns that the most relevant stimulus dimension is likely to occur in a particular spatial location across trials, the subject could begin orienting attention to this location in an anticipatory fashion before a stimulus is even revealed on trial $(i + 1)$.[1] To keep the notation for between- and within-trial dynamics separate, we use $\boldsymbol{\alpha}^{*}(t)$ to denote the within-trial attention vector at the *t*th moment in time, and we have dropped the superscript designating trial number for convenience. To initialize $\boldsymbol{\alpha}^{*}(t)$ on Trial *i*, we set $\alpha^{*}(t) = \alpha^{(i-1)}(t) / \sum_j \alpha_j^{(i-1)}(t)$, such that the within-trial module is initialized with normalized between-trial weights from the most recent update. We made this choice in order to *inform* orientation in the within-trial module using between-trial weights, but in a way that does not make strong assumptions about scale equivalence between feedback-based and search-related updates to attention.

Because it is central to our theory that between-trial learning and within-trial information sampling involve a common set of mechanisms, we needed a way to modify the feedback-based attention update mechanism in Equation 3 to account for the unsupervised aspect of within-trial dynamics. The between-trial module reduces the probability of future errors by redistributing the attention weights in a gradient-based manner, given the model's predicted response probabilities for each available option as well as the true category label. Because the true category label is not known by the observer until after the within-trial process terminates, however, we made the choice to calculate the within-trial attention update in reference to the model's current best guess about the true category label. Using Equation 8 as the specification for momentary evidence, we define a dynamic loss function as follows:

$$\mathbf{G}(t) = \nabla_{\alpha^{*}(t)} \{ \log [V^{(i)}_{\text{leading}}(t)] \}, \tag{9}$$

where $\mathbf{G}(t) = [g_1(t) \ \ldots \ g_D(t)]$ is the gradient-based update vector, and $V^{(i)}_{\text{leading}}(t)$ denotes maximum evidence across response options at moment *t*. Mirroring Equation 4, attention is updated at each timestep using the equation:

$$\alpha^{*}(t+1) = \alpha^{*}(t) + \gamma_w \mathbf{G}(t), \tag{10}$$

where γ_w is the within-trial learning rate. Because fixations are calculated directly from the attention update described by Equation 9, this specification supports confirmatory search behavior, such that attention will orient toward the dimensions that support further gains in evidence for the leading accumulator. We acknowledge that other, more balanced approaches could have been taken instead. For example, we could have calculated a separate *G* vector that maximizes evidence for each of the *C* candidate category labels. In this case, the update to attention in Equation 10 could have been the sum or a weighted combination of all *G* vectors, or we could have applied a maximum gain selection criterion similar to SEA. While alternative approaches merit additional investigation in future work, the selected implementation is consistent with observed confirmatory biases in human learning (Lefebvre et al., 2022; Nickerson, 1998; Talluri et al., 2018), has an advantage of computational parsimony over some unbiased alternatives, and was demonstrably sufficient for predicting human-like attention dynamics across the four simulation case studies that will be presented in sections to follow.

An important contribution of our work is to put forth a generative framework that makes explicit predictions for what dimensions will be fixated at each moment in time. To achieve this, we must consider two cases of dimension spatial arrangements. In the first case, stimulus dimensions are separated into different spatial locations (i.e., segregated dimensions). Here, eye-tracking measures provide direct measures of attention for specific dimensions, assuming only one spatial location can realistically be fixated at a time. In the second case, stimulus dimensions overlap in space (i.e., integrated dimensions). Fixation information therefore cannot distinguish between dimensions that are selectively attended (i.e., via covert attention), and dimensions that are ignored.

In the segregated case, we can identify the fixated dimension directly from the most recent update to the attention vector. Letting $L_j(t)$ denote the fixation index for each Dimension *j* at Time *t*, we can define the following equation:

$$L_j(t) = \begin{cases} 1 & \text{if } \max_j |g_j(t)| = |g_j(t)| \\ 0 & \text{otherwise} \end{cases}.$$

In other words, we assume that a fixation will be directed toward the dimension that is expected to provide the largest absolute amount of information, according to one's current representation.

In the integrated case, the specification is similar but additionally accounts for the fact that at least one subset of dimensions spatially overlap. Letting $\mathcal{S}$ denote the set of dimensions that have a spatial location that is identical to that of the most informative dimension, we can specify:

$$L_j(t) = \begin{cases} 1 & \text{if } \{ j \in \mathcal{S} : \max_j |g_j(t)| = |g_j(t)| \} \\ 0 & \text{otherwise} \end{cases}.$$

Hence, if one dimension is deemed to be the most informative at a given moment in time, then any dimensions that occupy the same location in space (i.e., shape and color dimensions of a particular item) will also be fixated within the same moment.

The distinction between the segregated and integrated cases exemplifies the fact that fixation data provide only minimal constraints on encoding. Although we assume that fixating to a dimension is a necessary condition for encoding a feature value, it is not guaranteed to be sufficient. In the case of selective attention, an observer could be overtly fixating on a particular spatial location, but only attending covertly to a subset of information contained therein

[1] It would be possible to define attention as a global optimization process for each individual trial; however, the presence of anticipatory attention orientation suggests to us that attention can be roughly approximated as a combination of single updates across trials, along with attention on each trial that is inherited from the between-trial dynamics. In other contexts, global definitions of the relevance of dimensions may prove more effective.

(Rutman et al., 2010). In these cases, it would be problematic to assume that fixation data alone give us complete and direct information about how time spent looking at a dimension relates to the probability of encoding. As such, we assume that encoding is based on a thresholded, cumulative sum of attention over the course of dwell time, such that:

$$Q_j(t) = \begin{cases} 1 & \text{if } \sum_t L_j(t)\alpha_j^*(t) \geq \kappa \\ 0 & \text{otherwise} \end{cases}.$$

Hence, when the cumulative attention applied to a fixated dimension exceeds the threshold κ, a feature is considered to be "encoded." After a trial terminates ($t = RT$), the information in $Q(RT)$ is stored in M^* so that only encoded feature values can be imputed on subsequent trials (Equation 7). As discussed in the Attention Is the Mechanism of Learning section, this specification is considerate of findings of improved memory for items that are covertly attended through endogenous means (Addleman et al., 2018; Botta et al., 2019; Foster et al., 2020). Recent neuroimaging work has suggested, however, that both endogenous and exogenous (i.e., related to salience) modes of covert attention facilitate perceptual encoding, but endogenous attention uniquely facilitates subsequent readout of visual information, such as what would be required for a category judgment (Dugue et al., 2020). While we do not account for the potential impact of exogenous covert attention on encoding probability here, this is a topic of future work that will be considered in the General Discussion.

Relating encoding dynamics back to the working representation of the stimulus probe as described in the Stimulus Encoding section, we again note that encoding a feature alters the distribution of values that are imputed into the working probe representation $\mathbf{e}^*(t)$. At the beginning of a trial when the feature value that occupies dimension j has not been encoded yet, $Q_j(t)$ will be equal to zero, and feature values r_h will be imputed into $\mathbf{e}_j^*(t)$ with probability $P(t)(e_j^{(i)}) = r_h)$ (Equation 6). After the cumulative attention applied to the spatial location of dimension j exceeds κ, $Q_j(t)$ is set to 1 to indicate that a feature was encoded. From that point on, the value $e_j^{(i)}$ will be imputed into the working representation $e_j^*(t)$ with probability $\theta Q_j(t)$, where θ represents encoding fidelity. With the remaining probability $1 - \theta Q_j(t)$, we assume that feature values will continue to be drawn from the distribution of previously observed values. Mathematically, we can write this process as a mixture of two probability mass functions, one containing a distribution $\pi_1(e_j^*(t) = r_h) = P(t)(e_j^{(i)} = r_h)$ over all expected feature values in $\mathcal{R}_j$, and one defining a Dirac delta distribution:

$$\pi_2(e_j^*(t) = r_h) = \begin{cases} 1 & \text{if } r_h = e_j^{(i)} \\ 0 & \text{if } r_h \neq e_j^{(i)} \end{cases},$$

where all probability mass is centered at the true representation $e_j^{(i)}$. Hence, we can write the mixture of these two distributions as follows:

$$e_j^*(t) \sim \begin{cases} \pi_1(e_j^*(t) = r_h) & \text{with prob. } 1 - \theta Q_j(t) \\ \pi_2(e_j^*(t) = r_h) & \text{with prob. } \theta Q_j(t) \end{cases}.$$

With this specification, the working representation of a probe dimension continues to be stochastic after the encoding threshold has been surpassed, but is biased in the direction of the true stimulus feature value with magnitude θ. This allows the observer to continue to represent feature values that have occurred within the broader task context with some probability, even after the true feature value of the stimulus is known. This appears to be important in situations where novel features or combinations of features are introduced (e.g., Case Studies 1 and 4), such that the observer is able to reorient if the fixated dimension provides information with unknown or contrary category information. For these types of situations, SEA contains a mechanism for a "sampling bonus," which can be artificially imposed to induce continued sampling. AARM's proposed way of balancing encoded stimulus information with available task feature information, however, naturally produces reorientation behavior in the presence of novel stimuli without modification.

A brief example of how attention reorients within a trial is shown in Figure 4E. Essentially, attention orients to the dimensions that have the best chance of resolving the conflict among the active choice options. In this case, the first encoded dimension (i.e., the food source) activated the chipmunk and squirrel categories, and so the deliberation now turns toward dimensions that accentuate differences between them. The next most important dimension is the tail. Attention, therefore, reorients to the tail dimension so that it will be subsequently fixated in Figure 4D. Additional elaboration of the attention updating process is provided in Figure 5 using the stimulus structure from Figure 1, as described in the Exemplar Activation section. In summary, the gradient update is initially maximized in Dimension D1, given that D1 is relevant to categorization across all exemplars ($t = 1$). After a D1 feature value is encoded for the current stimulus, exemplars from Categories "A1" and "A2" are selectively activated on the basis of similarity to the encoded information ($t = 2$ and $t = 3$). Attention then reorients to D2 in order to distinguish between the two most active categories, and encoding a D2 feature value facilitates retrieval of exemplars from Category "A2" ($t = 3$ and $t = 4$).

Summary

In this section, we provided the technical details of AARM as they relate to between- and within-trial dynamics. Although the notation can become complex when dealing with dynamics at two different time scales, the intuition of the model is far simpler. When provided with a stimulus (Figure 4A), observers sample information selectively in order to make an accurate and time-effective choice. With experience, observers learn to prioritize dimensions that help them separate stimuli into categories. When faced with a choice, observers deploy selective attention in a manner that is consistent with a learned prioritization map and begin encoding relevant stimulus features accordingly. The encoding process constructs a psychological representation of the stimulus probe (Figure 4B), which in turn activates memory traces of similar exemplars (Figure 4C). The retrieved exemplars are typically associated with a category label, such that the observer can accumulate evidence for a response (Figure 4F). For complex stimuli, these response options compete for selection and necessitate sampling of additional stimulus dimensions. Consequently, attention reorients to the dimensions that would facilitate a comparison among the most competitive options (Figure 4E) and can produce a shift in the fixated location (Figure 4D). This process continues until a decision threshold is

reached, based on the relative difference between the evidence among the response options.

In the next part of the article, we describe the results of four theoretical case studies that demonstrate how the between- and within-trial modules of AARM contribute to human-like predictions of choice and eye-tracking behavior across a comprehensive set of challenging scenarios. Case studies are divided into two sections to explicate the theoretical tenets of AARM: (a) humans seek out information about new stimuli in a manner that is influenced by their individual learning experiences and (b) attention is sensitive to hierarchically organized information, such that the current state of knowledge guides future information sampling.

This study was not preregistered. Model code will be made available upon publication at https://github.com/MbCN-lab. The data used in Case Study 1 will be available upon reasonable request. The data used in Case Study 2 were made freely available online by Meier and Blair (2013) at https://doi.org/10.1016/j.cognition.2012.09.014.

Experience-Based Representations

Behavioral and modeling work has indicated that humans tend to form representations based on a subset of the available dimensions (Lee, 2001; Shepard & Arabie, 1979; Sloutsky, 2003; Tversky, 1977; Ullman et al., 2002). Eye-tracking work has further shown that after sufficient training with the structure of a task, humans tend to fixate to only a few informative dimensions before making a response (Blair et al., 2009; Rehder & Hoffman, 2005a, 2005b), and features that are overtly fixated are more likely to be stored in memory for later use (Irwin, 1996; Loftus, 1985). In cases where multiple sources of information might be equally sufficient for correctly identifying the category label across trials (see Figure 3, for a hypothetical example), the extent to which features are encoded during training potentially impacts how the observer elects to sample information from stimuli encountered at test.

The relationship between the storage of individual memories and generalized category representations has garnered great interest, particularly in regard to the role of the hippocampus in episodic inference and concept formation (Bowman & Zeithamova, 2018; Mack et al., 2016, 2018; Schapiro et al., 2017). Several theoretical accounts have suggested that generalization does not require the formation of integrated concept representations, but rather can be achieved by the encoding and strategic retrieval of discrete memory traces (Hintzman, 1984; Kruschke, 1992; Kumaran & McClelland, 2012; R. Nosofsky, 1988). While some functional magnetic resonance imaging (fMRI) work has supported exemplar-based accounts by identifying similar hippocampal activation for both recognition and categorization judgments (Mack et al., 2013; N. Nosofsky et al., 2012), other work has provided evidence of associative inference functions of the hippocampus that arguably extend beyond item-specific memory storage. For example, repetition effects in the hippocampus have provided evidence that overlap between the current stimulus and existing memories impact how new items are encoded (Richter et al., 2016; Zeithamova et al., 2012, 2016; Zeithamova & Preston, 2017). In addition to encoding new information in reference to recent experiences, other work has provided evidence that humans make predictions about future events that are shaped by memories of the past (De Brigard et al., 2016; Van Hoeck et al., 2013), often recruiting the same networks that are involved in encoding and retrieval (De Brigard et al., 2013; De Brigard et al., 2015).

Given these insights on the impact of memory on category representations, Case Study 1 investigates memory-dependent sampling and decision dynamics in a category learning paradigm with multiple informative dimensions (Blanco & Sloutsky, 2019). In its original presentation by Galdo et al. (2021), the between-trial module of AARM is not equipped to account for variability in feature encoding over the course of learning. With the within-trial module, however, we can gain insight into how encoding variability related to selective attention during training might give rise to different patterns of information sampling behaviors in the presence of novel stimuli.

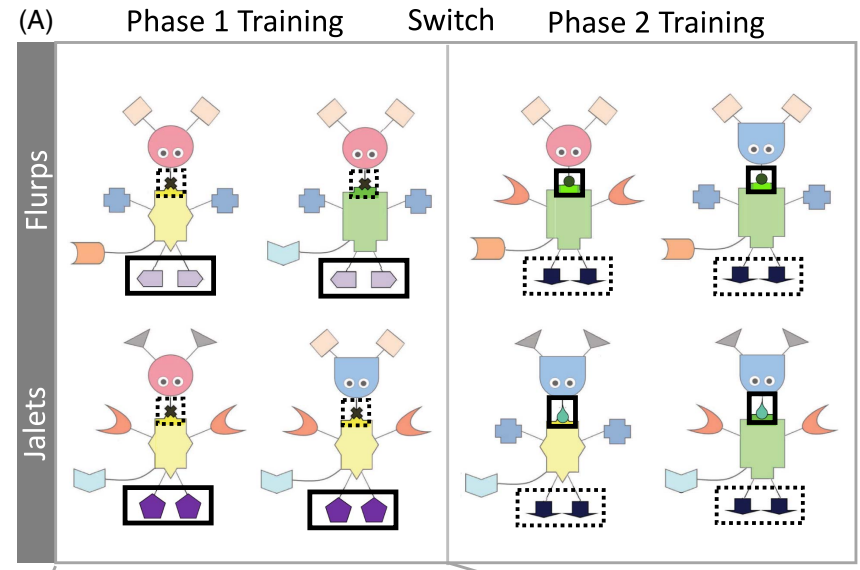
Case Study 1: Dimension Biases

To investigate the impact of attention-mediated encoding variability on subsequent information sampling and retrieval, we used a paradigm that was developed by Blanco and Sloutsky (2019). Here, we discuss behavioral and eye-tracking data that were collected from a cohort of 38 adults while they completed the same paradigm (Blanco et al., under review). The task paradigm will be summarized here, but the reader is directed to Blanco and Sloutsky (2019) for additional details.

As illustrated in Figure 6, categories were defined with a rule-plus-similarity structure, such that one "deterministic" dimension was perfectly predictive of category membership, and five "probabilistic" dimensions provided good but imperfect category information across trials (80% cue validity). An additional "irrelevant" dimension contained the same feature value across stimuli, and therefore contained no category-diagnostic information. Stimuli were images of alien-like characters that were composed of seven dimensions: antenna, head, body, button, hands, feet, and tail. Each dimension could take on one of a discrete set of features that varied on the basis of color and shape (i.e., the terminal ends of antennae could be either beige rectangles or gray triangles; hands could be either blue crosses or red half-moons, etc.). During the instructions, participants were informed that they would be seeing different creatures called Flurps and Jalets, and that their task was to figure out which species each creature belonged to. Participants also received instructions about the category structure. The features of each dimension were shown to participants in isolation, along with the message that "most" (for probabilistic dimensions) or "all" (for the deterministic dimension) creatures belonging to a particular category shared that feature. No information about the irrelevant dimension was provided during the instructions.

The task was divided into two phases that contained complementary sets of stimuli. Each stimulus in Phase 1 had a counterpart in Phase 2 that contained the identical configuration of probabilistic features, and was mapped to the same category label. The deterministic and irrelevant dimensions, however, switched roles between Phases 1 and 2. As shown in Figure 6A, for example, "feet" features that were deterministic in Phase 1 were replaced with a novel irrelevant feature in Phase 2, and the irrelevant "button" feature that occurred in all Phase 1 stimuli was replaced with one of two novel deterministic features in Phase 2. Participants were not informed that the switch would occur and did not receive any explicit instructions about the postswitch feature mappings.

Figure 6
Paradigm and Stimuli Used in Case Study 1



(B) Phase 1 Test

	Feature	D	I	P1	P2	P3	P4	P5	Feature	D	I	P1	P2	P3	P4	P5
Match Trials	Flurps	0	2	1	0	0	0	0	Conflict Trials: Jalet D, Flurp P	1	2	1	0	0	0	0
		0	2	0	1	0	0	0		1	2	0	1	0	0	0
		0	2	0	0	1	0	0		1	2	0	0	1	0	0
		0	2	0	0	0	1	0		1	2	0	0	0	1	0
		0	2	0	0	0	0	1		1	2	0	0	0	0	1
	Jalets	1	2	0	1	1	1	1	Conflict Trials: Flurp D, Jalet P	0	2	0	1	1	1	1
		1	2	1	0	1	1	1		0	2	1	0	1	1	1
		1	2	1	1	0	1	1		0	2	1	1	0	1	1
		1	2	1	1	1	0	1		0	2	1	1	1	0	1
		1	2	1	1	1	1	0		0	2	1	1	1	1	0

Note. (A) Illustration of stimuli, which participants were asked to sort into fictional "Flurp" and "Jalet" species types. Each stimulus contained seven dimensions (antennae, head, button, body, hands, feet, and tail). In Phase 1, one dimension (e.g., feet; outlined by solid box) was deterministic, one was irrelevant (e.g., button; outlined by dashed box), and five were probabilistic (all un-outlined features) in terms of cue validity. After an undisclosed "switch," the deterministic dimension from Phase 1 became irrelevant in Phase 2 and the irrelevant dimension became deterministic. (B) Characteristics of stimuli presented at test. Match items were drawn directly from the training set, such that deterministic and probabilistic dimensions carried the same feature-to-category mappings. Conflict items contained novel configurations of features, such that the deterministic and probabilistic dimensions carried opposite category mappings. In the table, unique feature values within each dimension are indicated by 0, 1, and 2. See the online article for the color version of this figure.

Each phase consisted of a training stage (with feedback), followed by a testing stage (without feedback). In the training stages, each of 10 unique items (5 from each category) from the relevant stimulus set were presented three times in random order (30 trials total). All stimuli were presented in the center of the screen, and each dimension occupied the same spatial location across trials. Participants made category responses by pressing buttons on a controller. After a response was made, participants were given corrective feedback. During Phase 1 training, feedback was very descriptive in an effort to encourage both attention to the overall appearance of

the stimulus as well as the deterministic dimension. For correct responses, feedback took the form of "Correct this is a Flurp. It looks like a Flurp and has the Flurp feet." Feedback following incorrect response took the form "Oops this is actually a Jalet. It looks like a Jalet and has the Jalet feet." Feedback during Phase 2 training was simplified so that participants were free to learn the postswitch feature-to-category mappings on their own. As such, feedback for a correct response was "Correct this is a Flurp," and feedback for an incorrect response was "Oops this is actually a Jalet."

Testing stages consisted of 10 trials from each of two conditions that were presented in random order (20 trials total). Participants responded to each item by selecting a category label, but no feedback was provided. Items in the "match" condition were identical to the stimuli presented during training. Items in the "conflict" condition contained novel configurations of previously encountered features. As shown in Figure 6B, each conflict item contained a feature in the deterministic dimension that was associated with one category label and features in the probabilistic dimensions that were associated with the opposite category label.

Continuous eye-tracking data were collected while participants completed the task using an EyeLink 1000 eye tracker at a sampling rate of 250 Hz. To preprocess the data, eight nonoverlapping rectangular areas of interests (AOIs) were defined surrounding the spatial locations of features in each dimension. Six out of seven dimensions occupied only one AOI, and the "hands" dimension occupied two. Fixation points were mapped to a particular dimension if they fell within the bounds of the relevant AOI, and were otherwise excluded from the analysis.

In the sections to follow, we consider the data from Blanco et al. (under review) in two parts. Case Study 1A uses data from the training and testing stages of Phase 1 to observe how information sampling during training relates to response probabilities in the presence of previously seen items (match) and novel configurations of features (conflict). Case Study 1B uses data from the transition between the testing stage of Phase 1 to the training stage of Phase 2 to observe how participants redistribute attention after the most reliable source of information suddenly becomes irrelevant for identifying category membership. In both sections, we provide simulation results from AARM alongside the observed data in order to demonstrate the model's ability to predict human-like information sampling and decision behaviors. Throughout, we will refer to deterministic, probabilistic, and irrelevant dimensions as "*D*," "*P*," and "*I*," respectively.

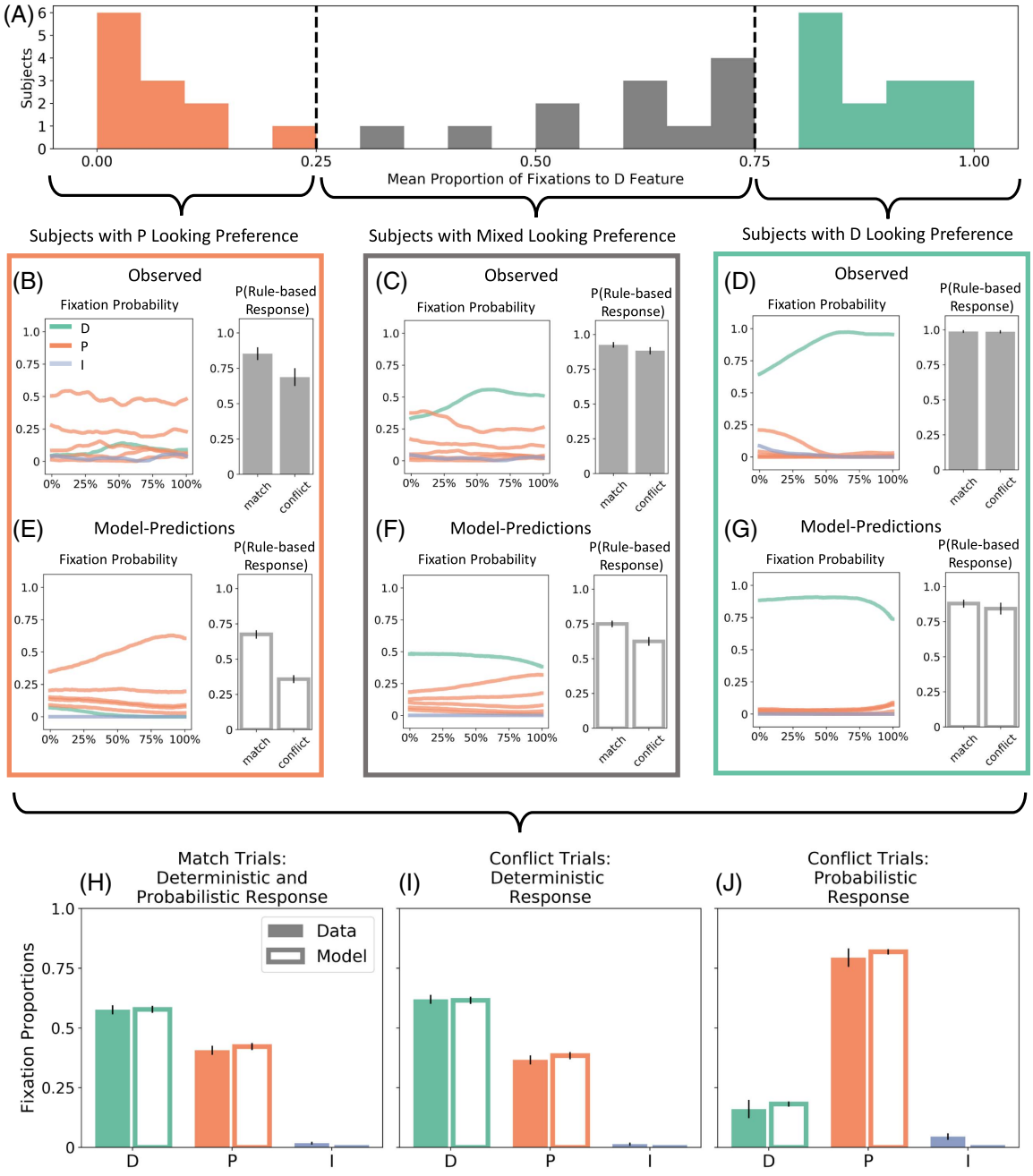
Case Study 1A: Conflicting Information

The within-trial module of AARM makes specific predictions about how feature encoding over the course of learning impacts information sampling behaviors and decision processes in the presence of new stimuli. The paradigm developed by Blanco and Sloutsky (2019) provides a unique opportunity to test these predictions, given that multiple dimensions provide information that is independently relevant to the task. Like the example provided in Figure 2, participants may achieve similarly high accuracy during training whether they selectively attend to the *D* dimension, a subset of *P* dimensions, or a combination of the two. In addition to the emergence of fixation preferences for particular dimensions, responses to test items in the current paradigm provide insight into how attention was distributed. Specifically, test items drawn from the conflict condition contain a combination of features that are associated with opposite category labels. Responses consistent with information in the *D* dimension (i.e., RB responses) could therefore be interpreted as evidence that the participant learned to selectively attend to that dimension during training. Similar logic holds for the *P* dimension as well, such that selective attention to any combination of *P* dimensions could manifest in P-consistent responses.

For our purposes, we performed analyses and simulations that were considerate of individual differences in fixation preferences at test. Because feedback was only provided during training, we considered fixations at test to be a stable indicator of postlearning attention. A histogram of fixation preferences across subjects is shown in Figure 7A. Considering only the first 10 test trials of Phase 1, we observed a relatively balanced distribution of fixation preferences for the *D* dimension as determined by normalized dwell times in the form $\frac{Dwell_D}{Dwell_D + Dwell_P}$ (mean = 0.541, min = 0.00, max = 1.00). We organized subjects into three groups on the basis of these fixation preferences: Group (1) looking preference for *P* dimensions ($\frac{Dwell_D}{Dwell_D + Dwell_P} \leq 0.25$; 12 subjects); Group (2) mixed looking preference ($0.25 < \frac{Dwell_D}{Dwell_D + Dwell_P} < 0.75$,12 subjects); and Group (3) looking preference for the *D* dimension ($\frac{Dwell_D}{Dwell_D + Dwell_P} \geq 0.75$,14 subjects). Panels B–D of Figure 7 show mean trajectories of fixations within Phase 1 test trials (line plots) and subsequent response proportions to match and conflict stimuli (bar plots) for Groups 1, 2, and 3, respectively. Within-trial fixations to each of the seven stimulus dimensions were calculated as percentages of the RT, binned by steps of 0.1%, averaged across trials, and smoothed using a moving window of size 1%. *P* dimensions were rank ordered within subject according to mean fixation probability across trials prior to aggregation in an effort to account for spurious differences among *P* dimension preferences when interpreting the results.

To simulate fixations and responses with AARM, we used a single set of parameters across subjects for the between-trial module (Table B2 in Appendix B). Though other approaches could have been taken, we made this choice in an effort to isolate individual differences in sampling trajectories to feature-level encoding fidelity, as used by the within-trial module. For each subject, we interpolated the number of encoded *D* features across unique training stimuli based on each subject's observed mean proportion of fixations to *D* at test (Figure 7A). Subject-level proportions of *D* fixations were split into quantiles and mapped to a discrete value *U* between 5 and 9 (inclusive) to represent the number of unique training trials out of a possible 10 during which the *D* feature was encoded (i.e., fewer fixations to *D* at test implied fewer *D* features were encoded during training). The matrix M^*, which contains the encoding status for the features of the stored exemplars (Equation 5), was then modified for each subject accordingly. A random selection of *U* elements in the column of M^* corresponding to the *D* dimension was set to 1 (meaning "encoded") and was otherwise set to 0 (meaning "unencoded"). All elements in M^* that corresponded to *P* and *I* dimensions were set to 1. We then simulated 1,000 match and conflict trials for each subject. The parameter values and initialized attention weights used for our simulations were otherwise fixed across subjects and task conditions and were optimized with respect to the observed patterns of data in aggregate. Because no feedback was given at test, trials were simulated in

Figure 7
Case Study 1A: Conflicting Information



Note. CIs = confidence intervals; RB = rule-based; AARM = adaptive attention representation model. (A) Subject-level mean proportions of fixations to the deterministic dimension. (B–D) Left panels show mean proportions of fixations (*y*-axis) to each of the seven dimensions through time (*x*-axis). Right panels show observed means and 95% CIs of proportions of RB (i.e., responses consistent with the deterministic feature) responses across match and conflict trials. (E and F) Fixation and response data were generated using AARM's within-trial module. (H–J) Mean proportions of fixations to the deterministic, probabilistic, and irrelevant dimensions across observed (filled bars) and model-generated (unfilled bars) trials, collapsed across groups. (H) Fixation proportions across match trials. (I) Fixation proportions across conflict trials on which responses were consistent with the deterministic dimension. (J) Fixation proportions across conflict trials on which responses were consistent with the majority of probabilistic dimensions. See the online article for the color version of this figure.

isolation without subsequent updating of the stored exemplars in matrix *X*. Parameter values used for our simulations are provided in Table B2 in Appendix B.

As shown in Figure 7E–G, AARM's predictions reflect several important elements of the observed data. By simply accounting for encoding variability for feature values observed in the *D* dimension, AARM was able to predict the observed effect of increasing proportions of fixations to *D* from Group 1 to Group 3. Considering response probabilities, AARM further predicted the observed effect of increasing proportions of RB responses to conflict trials from Group 1 to Group 3. Although there are discrepancies between the observed data and the simulations in terms of, for example, the probabilities of initial fixations and mean response probabilities within each group, we consider these results to be promising overall. While we intentionally relegated individual differences in test behavior to encoding efficiency and held all other mechanisms constant, future work will more thoroughly investigate how partial encoding of individual items impacts the trajectory of between-trial learning in addition to within-trial sampling.

For our current purposes of qualitatively assessing AARM's theoretical assumptions about within-trial dynamics, Figure 7H–J provides a proof-of-concept. Proportions of observed and simulated (same simulations shown in Panels E–G) fixations to *D*, *P*, and *I* dimensions were averaged across subjects within each test condition of Phase 1. Panel H shows that correct responses to Match stimuli (*D* and *P* features have matching category mappings) were preceded by a slight fixation preference for *D* compared to *P* when the data are considered in aggregate. Because *D* and *P* dimensions carried opposite category mappings on conflict trials, we back-sorted the data by response in order to observe potential differences in fixation probabilities. Panel I shows that responses consistent with the *D* feature-to-category mapping were preceded by a fixation preference for *D* over *P* dimensions. Panel J shows the opposite fixation bias, such that responses consistent with the *P* feature-to-category mapping were preceded by a fixation preference for the *P* dimensions over *D*. As shown by the unfilled bars in Panels H–J, AARM predicts patterns of response-dependent fixations that closely match what we see in the data.

Here, we demonstrated AARM's predictions about attention allocation and decision-making as a result of successful learning and encoding efficacy. In the decision-making literature, the widely supported integrate-to-bound perspective suggests that information is sampled from multiple sources of information within a trial, and choices are made when the cumulative evidence in favor of one option exceeds a predetermined threshold (e.g., Ratcliff, 1978). Other work used eye-tracking methods to show that proportions of fixations to two competing options in value- and preference-based decision tasks are directly related to choice probability (Krajbich & Rangel, 2011; Krajbich et al., 2010; S. Smith & Krajbich, 2019a, 2019b; Thomas et al., 2019). Extending this logic to categorization decisions, the intuition is simple: If an observer does not look at a particular dimension during a trial, the unseen feature will not contribute to the choice. While models that conceptualize attention as trial-level weights can adequately predict average proportions of responses in a variety of cases, the relationship between attention weights and information sampling behaviors has remained underexplored. AARM, however, makes specific predictions about which stimulus dimensions will be prioritized, attended, and sampled within a trial, and how sampling affects subsequent responses in the presence of new stimuli. Our simulations show that AARM not only predicts the same contingency between fixations and responses that are observed across experimental conditions of information consistency (Figure 7H–J), but can also predict individual differences in dimension prioritization as a result of encoding individual features (Figure 7E–G).

Case Study 1B: Shifting Information Relevance

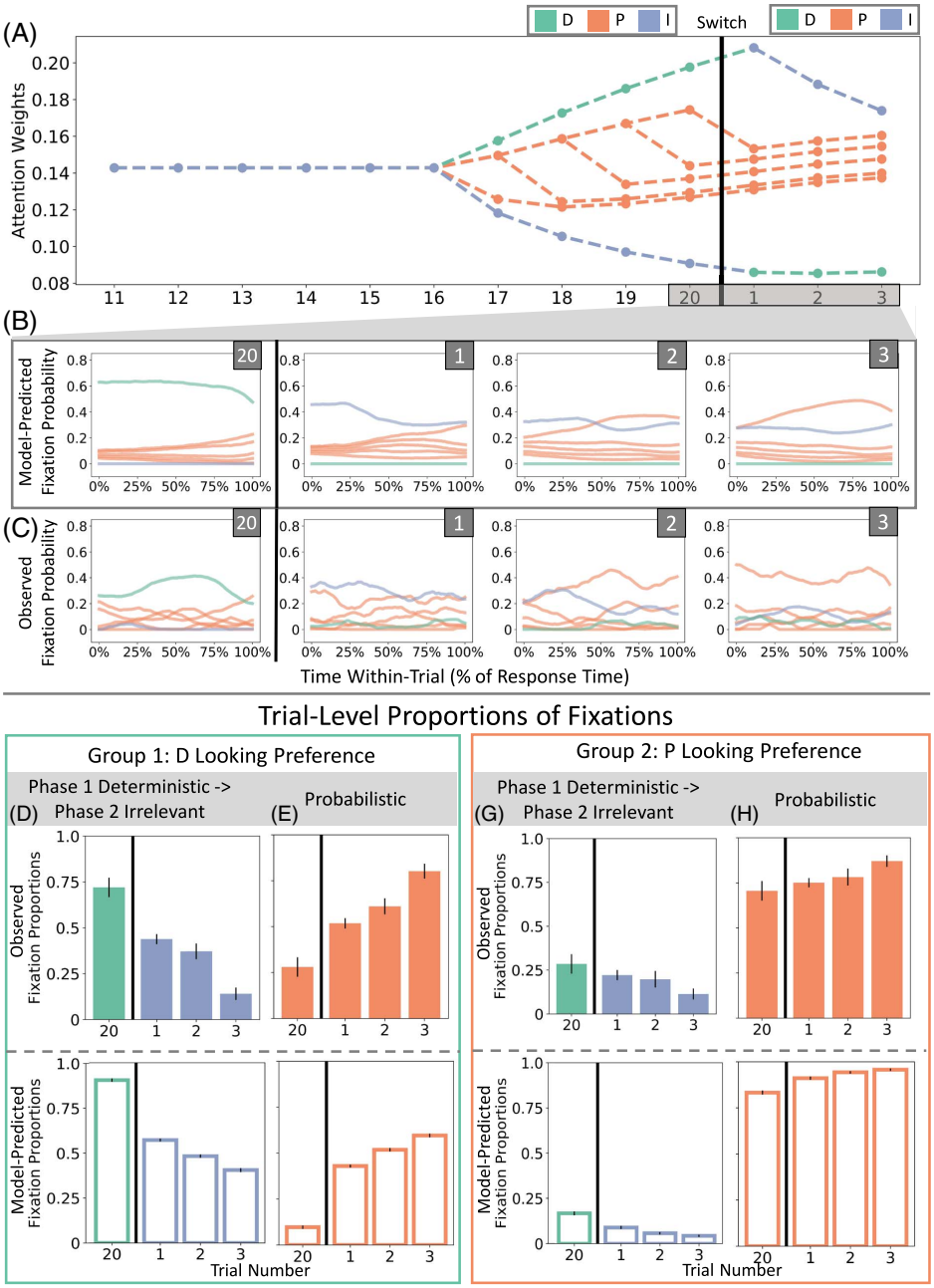
We next explored how within- and between-trial attention dynamics interact over the course of learning. Previous work has demonstrated that adult learners use selective attention to prioritize the most relevant information (e.g., Desimone & Duncan, 1995) and can adapt to changing categorization rules via set shifting (e.g., Chiu & Yantis, 2009). Although engaging selective attention can lead to faster, more efficient categorization, it can also result in learned inattention (Hoffman & Rehder, 2010). When an observer ignores a dimension after learning that it is uninformative for the task, it is often difficult for the observer to identify if and when the ignored feature becomes relevant at some point in the future. The Blanco and Sloutsky's (2019) paradigm provides an opportunity to observe how learners adapt to abrupt changes in information relevance when the *D* and *I* dimensions from Phase 1 switch roles during the transition to Phase 2. With the addition of eye-tracking data (Blanco et al., under review), we gain insight into how selective attention and feature encoding modulate the impact of the switch on information sampling. Here, we discuss the observed effects of the switch on information sampling behaviors, and how these effects are explained by AARM. For clarity, we refer to the dimension that was deterministic in Phase 1 and irrelevant in Phase 2 as "*D*/*I*," and we refer to its counterpart as "*I*/*D*." The observed and model-predicted fixation results that are relevant to the current discussion are shown in Figure 8. Note that all panels in Figure 8 show data from the final trial in Phase 1 (Trial 20 of the Phase 1 test stage) and the first three trials in Phase 2, separated by a vertical black line.

We first discuss the observed results, as shown in Figure 8C, D–H. Although participants were not informed of the switch, the aggregate data shown in Figure 8C indicate that participants quickly realized that the dimension that was most reliable for identifying category membership in Phase 1 (left-most panel; green line) was no longer reliable in Phase 2 (remaining panels; purple line). Across the four trials of interest, we observe a steady decrease in the proportion of fixations to the *D*/*I* dimension. This awareness, however, did not extend to the change in relevance for the formerly irrelevant dimension: Participants continued to ignore *I*/*D* in Phase 2, presumably as a result of learned inattention incurred during Phase 1 (Hoffman & Rehder, 2010). Instead, participants reoriented attention to a *P* dimension after the switch. Prioritization of *P* and inattention to *I*/*D* persisted across the entirety of Phase 2, beyond the initial three training trials shown in Figure 8 (mean proportion of fixations across Phase 2: $I/D = 0.168$, $P = 0.629$, $D/I = 0.203$).

In Case Study 1A, we observed how dimension-level fixation preferences related to choice behavior during conflict trials. We expected similar effects in the current case study, such that participants who tended to fixate to *D* during Phase 1 would demonstrate larger effects of attention reorientation after the switch to Phase 2. We specified two groups on the basis of proportions of *D*/*I* fixations across the last ten test trials of Phase 1. This differs

Figure 8

Case Study 1B: Shifting Information Relevance



Note. The final trial of Phase 1 and the first three trials of Phase 2 are of primary interest. In all panels, the vertical black bar represents the "switch" from Phase 1 (left) to Phase 2 (right). (A) Between-trial module-generated attention weights (points) for unique stimulus configurations. (B) 100 sequences of within-trial fixation and decision behaviors were generated by the within-trial module, using the specific sequence of stimulus configurations that each participant experienced. (C) Within-trial probabilities of fixating to each dimension were aggregated across subjects and plotted as a function of the percentage of observed response time. (D–H) Data and simulations for two groups, specified according to the proportion of fixations to D dimensions during the latter 10 trials of Phase 1 test. Group 1 showed a looking preference for D, whereas Group 2 showed a looking preference for P. Probabilities of fixating to the deterministic (D and G), or any of the five probabilistic (E and H) dimensions were averaged across observed (filled bars) and model-generated (unfilled bars) sequences. See the online article for the color version of this figure.

from the group delineations in Case Study 1A (proportions of *D* fixations across the *first* ten test trials of Phase 1) in order to be considerate of potential effects of the conflict trials on information sampling. Here, we specify Group 1: Looking preference for *D*/*I* ($\frac{Dwell_D}{Dwell_D+Dwell_P} \geq 0.75$); and Group 2: Looking preference for *P* ($\frac{Dwell_D}{Dwell_D+Dwell_P} \leq 0.25$). As shown in Figure 8D–E, Group 1 showed rapid deprioritization of *D*/*I* across the trials of interest, such that the mean proportion of fixations to *D*/*I* dropped from 0.72 to 0.14. This was accompanied by increased prioritization of *P*, such that the mean proportion of fixations across *P* dimensions rose from 0.26 to 0.78. Switch effects were substantially less severe in Group 2, as shown in the left panels of Figure 8G–H. The mean proportion of fixations to *D*/*I* dropped from 0.26 to 0.12 over the relevant trials, and the mean proportion of fixations to the *P* dimensions increased slightly from 0.72 to 0.81 over the same period.

To simulate data in AARM, we first used the between-trial module to determine a set of initialization weights for the four trials of interest. As in Case Study 1A, we made the decision to relegate differences in sampling behavior to mechanisms for feature prediction in the within-trial module. As such, the set of weights from the between-trial module was generated with a single set of parameters and was therefore constant across all subject-level simulations. Figure 8A shows the progression of attention weights generated by the between-trial updating mechanism in AARM, given the 10 unique training stimuli in Phase 1 and three training trials in Phase 2. Over the course of Phase 1, the model learns to increase attention to *D*/*I* and decrease attention to *I*/*D*. Following the switch, attention to *D*/*I* decreases and is redistributed among the *P* dimensions, while attention to the *I*/*D* dimension remains consistently low.

For within-trial simulations, each subject's proportion of *D*/*I* fixations during the last 10 test trials of Phase 1 was mapped to a discrete value between 6 and 9 (inclusive) to represent the number of training trials out of a possible 10 on which the *D*/*I* feature was encoded (i.e., fewer fixations to *D* at test implied fewer *D* features were encoded during Phase 1). The feature encoding matrix, M^*, was then modified as described in Case Study 1A. We probed the within-trial module with the exact sequence of stimuli that each participant actually observed; specifically, the last test trial of Phase 1 and the first three training trials of Phase 2. For each simulated trial, the model outputs were as follows: (a) a category response; (b) a response time equal to the number of iterations between initialization and self-termination; (c) a vector of predicted fixations with length equal to the RT, in which each element corresponded to a discrete dimension; and (d) a binary vector *Q*, which indicated whether each dimension was encoded ($Q_j(RT) = 1$) or not ($Q_j(RT) = 0$) by the end of the trial. After each trial in a sequence, the feature identity of the probe (e_i) was added to the matrix of stored exemplars (*X*) along with the corrective category feedback received by the participant. Similarly, *Q* was added to the exemplar encoding matrix, M^*. Elements that were set to 0 in M^* functioned as a mask over the corresponding feature values in *X*, such that feature information about a stimulus in a sequence was only accessible to the model on subsequent trials if it was encoded during training. We simulated 1,000 sequences of the four relevant trials for each subject. All simulations used the same set of parameters, which were optimized with respect to the observed data in aggregate (Table B2 in Appendix B). Figure 8B shows the model-predicted timecourse of fixation probabilities within each trial, averaged across subjects. At the end of Phase 1, the model predicts initial orientation to *D*/*I* on the basis of learned relevance (green line), but gradually reorients to a *P* dimension upon observing that the novel feature value in *D*/*I* is no longer relevant in Phase 2 (purple line). In Figure 8D–H, we observe that the model additionally predicts stronger effects of *D*/*I*-deprioritization and corresponding prioritization of *P* in Group 1 compared to Group 2. Despite minor qualitative discrepancies considering the precise timecourse of fixation probabilities, we consider these effects to be consistent with the observed data. With AARM's specification, the necessity of redistributing attention after the switch is contingent upon the extent to which the observer encoded features in the *D*/*I* dimension to begin with.

Hierarchical Category Structures

For decades, it has been known that hierarchical structures play an important role in guiding goal-directed behaviors, such that humans instinctively use superordinate sources of information to determine appropriate actions (e.g., Estes, 1972; Lashley, 1951; Miller et al., 1960). In task-cueing and task-switching paradigms, for example, humans are able to engage in different sets of RB behaviors in response to a stimulus-independent indicator (see Monsell, 2003, for review). In work by Meiran (1996), participants learned to classify digit stimuli as either odd/even or high/low depending on the shape or color of a background cue. While the authors observed notable switch costs, such that participants were slower to respond on the first trial after a rule switch, participants were indeed able to learn the mapping between background cues and the current rule in both predictable and unpredictable conditions of subtask sequences (see Allport et al., 1994; Rogers & Monsell, 1995, for similar results). Given that digit stimuli were drawn from a common distribution across odd/even and high/low subtasks, one interpretation of Meiran's (1996) results is that the background cue served as a hierarchically superordinate indicator of whether participants should attend to the digit's parity or magnitude on each trial.

Other work has suggested that humans use sources of contextual information to determine how to selectively allocate attention as well (e.g., Chun & Jiang, 1998; Chun & Turk-Browne, 2007; Crump et al., 2018; Egner, 2008; Vecera et al., 2014). In a contextual-cueing task conducted by Chun and Jiang (1998), for example, participants were able to use the global arrangement of stimuli as a cue for identifying the spatial location of a visual search target. These results are in line with seminal theories of memory and attention, in which contextual cues are bound to stimuli during encoding, and influence automatic attentional processing at test (Norman, 1968; Norman & Shallice, 1986; Shiffrin & Schneider, 1977).

In the three case studies to follow, we will use AARM to explore how hierarchical category structures give rise to distinct patterns of information sampling behaviors and within-trial changes in selective attention. We will first discuss results originally reported by Blair et al. (2009), which showed that humans prioritize information in a manner that is consistent with hierarchically organized stimulus dimensions. We will then expand the concept of hierarchical structures to environmental context as a superordinate dimension for determining how to appropriately distribute attention across the dimensions of the stimulus itself.

Case Study 2: Dimension Prioritization

In Case Studies 1A and 1B, we discussed how learning and memory impact the way humans decide which dimensions to sample when provided with a new stimulus. Case Study 2 further considers how the feature information contained within the current stimulus can impact the path and timecourse of attention allocation. As shown in Figure 1C, for example, hierarchically organized category structures contain jointly deterministic dimensions, such that the feature value contained in the superordinate (green squares) dimension indicates which of the available subordinate dimensions (orange triangles or purple crosses) is relevant to category membership. Such structures are ideal for studying within-trial dynamics, as they give rise to distinct temporal ordering effects of dimension-sampling behaviors between stimuli.

First, we present eye-tracking data from 41 subjects that were provided freely online by Meier and Blair (2013). The "1:1 condition" (equal frequency across category exemplars during training; Experiment 2) used the same stimuli and study design that were originally developed by Blair et al. (2009). Stimuli were fictional microorganisms, each containing a triad of equally spaced dimensions (organelles). Each dimension could take on one of two possible features, resulting in eight unique stimulus configurations. Stimuli were assigned to four categories (A1, A2, B1, and B2) based on a hierarchical category structure, such that one dimension (D1) indicated membership in an A or a B category, the second dimension (D2) differentiated between A1 and A2, and the third dimension (D3) differentiated between B1 and B2 (i.e., Figure 1). Participants completed 480 trials with feedback and were excluded from further analyses if they failed to exceed an accuracy criterion of 80% within the latter 120 trials (10 subjects; Meier and Blair, 2013). Following the analyses presented by the original study, we aggregated fixations to each dimension separately across Categories A and B items in the final 72 trials of the experiment. Mean fixation probabilities shown in Figure 9A, reflect striking differences in dimension prioritization between trial types. Replicating the findings from Blair et al. (2009), the results from Meier and Blair (2013) show that participants tended to fixate to the superordinate dimension first, then shift their gaze to the subordinate dimension that was relevant to the current trial while ignoring the alternative.

For our simulations, we first used the between-trial module to generate postlearning initialization weights after a single exposure to all eight unique stimulus configurations. After normalization (see the Attention Orientation section) D1 received the highest weight (0.505) and D2 and D3 each received lower but equivalent weights (0.248). We then used the within-trial module to simulate 1,000 isolated trials without feedback using A- and B-labeled probes as inputs. Outputs of each simulated trial were as follows: (a) a category response; (b) an RT equal to the number of iterations between initialization and self-termination; and (c) a vector of predicted fixations with length equal to the RT, in which each element corresponded to a discrete dimension (i.e., D1, D2, or D3). For generating dwell times, iteration units were converted to milliseconds by simple scalar multiplication. The simulated paths of fixations generated by AARM shown in Figure 9B closely resemble the observed behavior shown in Figure 9A. Across A and B probes, the model predicts a fixation bias toward the superordinate D1 dimension for the first 30%–40% of the trial's full duration before reorienting to the relevant D2 or D3 dimension. As a reconfiguration of the results shown in Figure 9A–B, Figure 9C shows total dwell times to each dimension, averaged across trials. Both observed (filled bars) and model-predicted (unfilled bars) results show approximately equal dwell times (600–650 ms) to the two dimensions that were relevant to each trial (D1 and D2 for Category A stimuli; D1 and D3 for Category B stimuli) and substantially shorter dwell times to the irrelevant dimension (D3 for Category A stimuli; D2 for Category B stimuli). Blair et al.'s (2009) original study provided compelling evidence that humans allocate attention in way that (a) favors features that are relevant within the current trial and (b) is observable via within-trial gaze fixation paths. Until recently (see Braunlich & Love, 2021), however, category learning models have not been subjected to constraints related to temporal ordering of prioritized dimensions. By explicitly defining how mechanisms for between-trial attention weights manifest in distinct paths of information sampling, AARM demonstrates the unique ability to relate latent theoretical constructs of attention to observable timecourses of within-trial behavior.

Case Study 3: Task Cueing

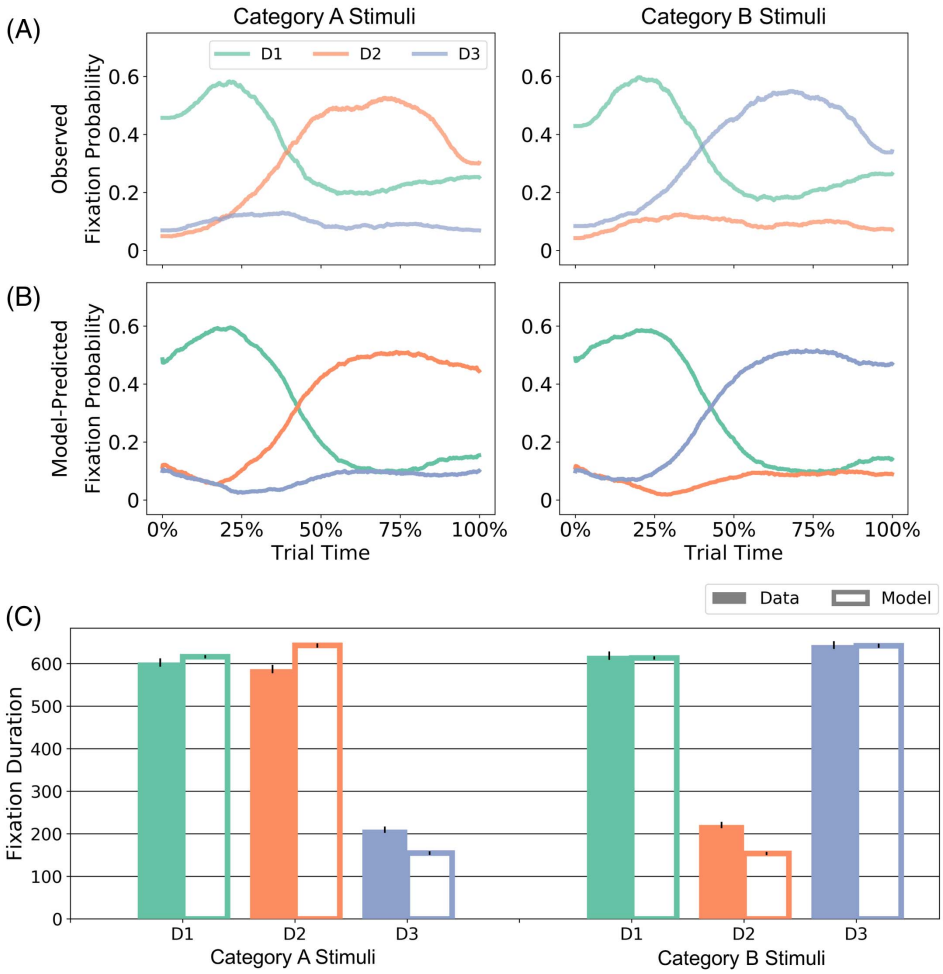
Here, we extend the concept of hierarchical structures to contextual cues as a superordinate dimension. Specifically, we used AARM to simulate response data across a set of RB and information-integration (II) subtasks in which a context dimension (i.e., background color) indicated which stimulus information was relevant for categorizing a common set of stimuli. While two-dimensional RB tasks require observers to categorize stimuli on the basis of a single dimension, II tasks require integration of feature information across multiple dimensions (Ashby et al., 1998; Maddox & Ashby, 2004; D. Smith et al., 2012). Humans and nonhuman primates have demonstrated an ability to learn both RB and II tasks, but are notably faster and more accurate at learning in the former case (Maddox & Ashby, 2004; D. Smith et al., 2012). Even when stimulus dimensions co-occur in space, behavioral evidence suggests that humans can selectively attend to task-relevant dimensions while ignoring the others (Ashby & Maddox, 2005).

We designed a simulation paradigm after O'Donoghue et al. (2020) to provide what we consider to be a clear demonstration of context-dependent learning and selective attention with AARM. However, we do not draw comparisons to observed data in the current case study. Given that data from O'Donoghue et al. (2020) were collected from pigeons rather than humans in an effort to study behavior in the absence of an analytic category learning system, we do not expect a model of human category learning like AARM to produce analogous behavioral results. Instead, the goal of Case Study 3 is to demonstrate AARM's ability to extend to categorization problems with (a) continuously valued dimensions; (b) spatially co-occurring dimensions; and (c) multiple levels of complexity (where II trials are assumed to be more complex than RB).

Figure 10A illustrates the hypothetical paradigm used here. Each point represents a combination of frequency (x-axis) and tilt angle (y-axis) feature values for a single Gabor patch that was created from a common, normally distributed stimulus space across trials. Contexts 1 and 2 denote RB subtasks, such that category membership was determined by high versus low frequency or large versus small tilt angle, respectively. Contexts 3 and 4 denote II subtasks, such that

Figure 9

Case Study 2: Hierarchical Category Structures



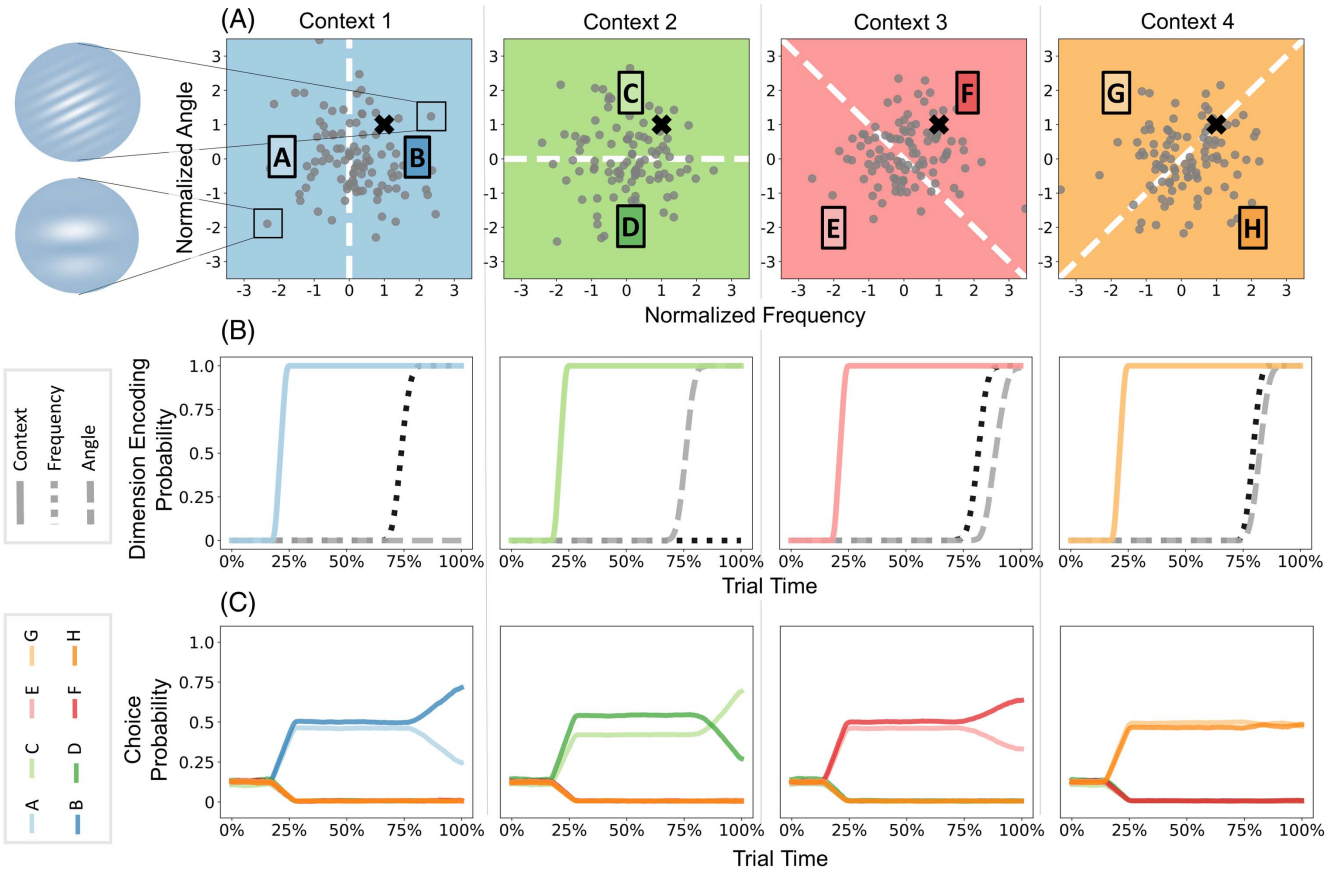
Note. AARM= adaptive attention representation model; CIs = confidence intervals. (A) Observed fixation data from Experiment 2 (1:1 condition) from Meier and Blair (2013) while participants categorized stimuli that belonged to categorized A (left panel) and B (right panel). (B) Within-trial fixation predictions generated by AARM aggregated across 1,000 probes from Categories A (left panel) and B (right panel). (C) Means and 95% CIs of dwell times to each stimulus dimension in milliseconds, calculated across Category A (left set of bars) and B (right set of bars) trials. Filled bars show observed data and unfilled bars show model predictions. See the online article for the color version of this figure.

category membership was determined by the integration of both frequency and tilt angle information.

To represent Gabor patches that varied continuously on frequency and tilt angle dimensions, we randomly drew 400 points (X, Y) from a bivariate normal distribution with means of 0 and standard deviations of 1. One hundred points were randomly allocated to each of the four contexts, and category labels were assigned according to the relevant rule as follows (Table 3).

Training stimuli with category labels matrix X, with the first two elements taking on continuous values, and the third element taking on a discrete value between 1 and 4 (inclusive) to represent context. All 400 training stimuli were iteratively introduced to AARM's between-trial module, and the posttraining weights for each dimension were as follows: 0.195 (X; frequency), 0.194 (Y; tilt angle), and 0.611 (context). As in the previous case studies, attention weights in the within-trial module were initialized to the posttraining values determined by the between-trial module. Four probes (e_i) were each introduced to the within-trial module 1,000 times without feedback. Feature values corresponding to the X and Y dimensions were set to 1 across probes (shown as crosses in Figure 10A), and context feature values corresponded to each of the four unique contexts in the task. By contrast to the previous case studies in which dimensions were spatially segregated, encoding probability was not gated by the output of the error gradient (i.e., gaze fixations). Instead, both attention weights and encoding probability were continuously updated throughout the trial for all dimensions simultaneously, given that context, tilt, and angle dimensions co-occurred in space. Outputs of each simulation were as follows: (a) a category choice

Figure 10
Case Study 3: Task Cueing



Note. RB = rule-based. II = information integration. (A) Illustration of stimuli and category delineations from four subtasks of a hypothetical experiment. Points represent Gabor patch stimuli that each take on a frequency value (*x*-axis) and a tilt angle value (*y*-axis). Background colors served as an indicator of the categorization rule for each subtask: frequency distinguished between Categories A and B in Context 1 (RB), angle distinguished between Categories C and D in Context 2 (RB) both frequency and angle were necessary for distinguishing between Categories E and F in Context 3, and Categories G and H in Context 4 (II). (B) Mean encoding probabilities of each dimension (*y*-axis) are plotted as a function of the percentage of time in between stimulus onset and response (*x*-axis). Solid, dotted, and dashed lines represent context, frequency, and angle dimensions, respectively. (C) Probabilities of making an A–H response (*y*-axis) plotted as a function of the percentage of time within trial between stimulus onset and response (*x*-axis). Each color represents an available category label, as shown in Panel A. See the online article for the color version of this figure.

(A–H); (b) a matrix of choice probabilities across categories at each timestep prior to self-termination; (c) an RT (number of iterations); and (d) a binary matrix indicating whether each dimension was encoded.

Table 3
Categorization Rules for Case Study 3

Context	Rule	Category
1	$X \geq 0$	*A*
	$X < 0$	*B*
2	$Y \geq 0$	*C*
	$Y < 0$	*D*
3	$-Y \geq X$	*E*
	$-Y < X$	*F*
4	$Y \geq X$	*G*
	$Y < X$	*H*

Within-trial averages of dimension encoding and choice probability across simulated trials are shown in Figure 10. As a reflection of the inherently hierarchical structure of the paradigm, we observe that context is prioritized across both RB (Contexts 1 and 2) and II (Contexts 3 and 4) subtasks, as illustrated by consistently early encoding of the context dimension (approximately 25% of the response time; Panel B). Analogs of selective attention emerge, however, when we observe which stimulus dimensions were encoded in each context. In accordance with the RB category structure learned in Context 1, the model tended to encode the frequency (X) dimension but not the tilt angle (Y) dimension across probes. By contrast, the model encoded tilt angle but not frequency when presented with a probe in Context 2. Humans are known to engage similar selective attention processes in the presence of integrated dimensions (see van Moorselaar & Slagter, 2020, for recent review), and these effects are accompanied by reduced subsequent memory for

task-irrelevant stimulus features (e.g., Olivers et al., 2011; van Moorselaar et al., 2014).

While context and only one stimulus dimension were encoded in RB Contexts 1 and 2, both frequency and tilt angle dimensions were encoded when probes were presented in II Contexts 3 and 4. This behavior is appropriate given the demands of the two II subtasks, in which integration of both stimulus dimensions is required to correctly identify category membership. Given probes with identical frequency and tilt angle feature values, AARM predicts maximum probabilities of correctly responding "B," "C," and "F" when presented in Contexts 1, 2, and 3, respectively. Because the stimulus probe is located on the category boundary in Context 4, AARM predicted an equal probability of selecting Categories G or H at the time of the response. Overall, the timecourses of choice probability predicted by AARM indicate both successful mapping of context to relevant candidate responses, as well as successful adoption of learned RB and II categorization rules.

The current case study is similar in scope to Case Study 2 in that context serves as a superordinate hierarchical indicator of the rule. Case Study 3 was meant to build upon the results of Case Study 2 in two important ways: (a) the RB versus II distinction demonstrates that AARM encodes spatially co-occurring dimensions independently or simultaneously, depending on the demands of the task; and (b) the use of continuous dimensions demonstrates that AARM can generalize learned information to categorization decisions about novel stimuli. Although the exact combination of probe feature values introduced at test was never observed during training, AARM was still able to use information about context to engage selective attention (Figure 10B) and predict responses consistent with learned RB and II subtasks (Figure 10C). Building upon the simulation results presented here, Case Study 4 will investigate how the nature of the learning environment can potentially modulate attention to context during training, and subsequent context-dependent behavior at test.

Case Study 4: Incidental Context

Empirical evidence has suggested that environmental context plays a role in memory encoding and retrieval even when the context is not directly relevant to the goals of the task (i.e., Godden& Baddeley, 1975; S. Smith et al., 1978). These effects extend from complex place to simple computerized manipulations of context (i.e., background color: Dulsky, 1935; Isarida and Isarida, 2007; Murnane et al., 1999; screen location: Dix and Aggleton, 1999; font size: Perfect, 1996). In the memory literature, these two types of context have been characterized as "local" and "global" context, respectively (Baddeley, 1982; Eich, 1985; Murnane et al., 1999). Whereas local context is associated with a subset of items and influences the representation of stimuli during encoding, global context refers to aspects of the learning environment that are independent of the to-be-remembered information (Hockley, 2008). Building from our examination of local context in Case Study 3, Case Study 4 tests the extent to which AARM can predict context-related differences in behavior, even if context is not an independently relevant dimension during learning.

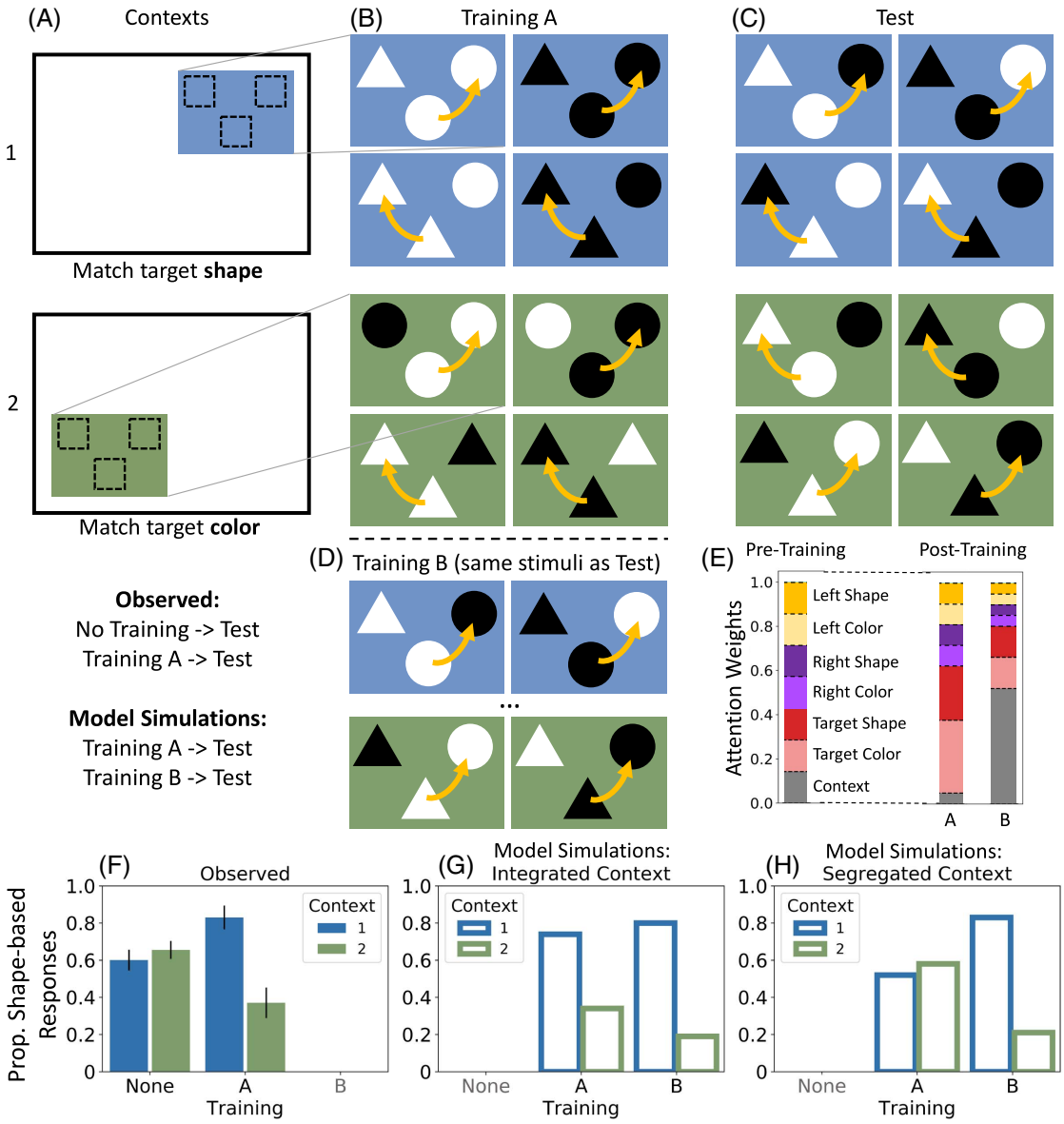
Case Study 4A: Context Integration

In the current case study, we focus on a paradigm that was developed by Sloutsky and Fisher (2008) to examine context-dependent generalization of learned concepts in 4- to 5-year-old children. Global contextual features (i.e., the color of a background rectangle and the stimulus's location on the computer screen; Figure 11A) co-occurred with categorization rules during training but were not independently relevant to the task. In the presence of novel test items with conflicting stimulus features, however, observed response biases indicated that participants had indeed learned the contingencies between contexts and categorization rules. Simulations with AARM provide an explanation for how the observed pattern of results might occur.

Stimuli were triads of items, with each item varying on the basis of shape (circle or triangle) and color (red or blue). Triads consisted of a target and two choice options, and the task on each trial was to select the choice option that matched the target on either the shape or the color dimension. Forty-two participants underwent training in which they responded to two types of trials with feedback: (a) in Context 1, all items in a triad had the same color, and participants had to respond on the basis of shape (Figure 11B, top: yellow arrows show paths from target to correct response); and (b) in Context 2, items had the same shape, and participants had to respond on the basis of color (Figure 11B, bottom). After 48 training trials, participants completed 16 test trials without feedback. By contrast to the training phase in which triad items only varied on a single dimension per trial, test triads were ambiguous. As shown in Figure 11C, one choice option matched the target on the basis of shape (and mismatched on color) while the other matched the target on the basis of color (and mismatched on shape). Half of the participants completed the test phase in Context 1, and the other half in Context 2. As a point of comparison, a separate group of 32 participants completed the test phase in both contexts after receiving no training at all. The behavioral results of the test phase are shown in Figure 11F. When tested in Context 1, participants who underwent training were more likely to respond on the basis of *shape* than the untrained participants. When tested in Context 2, trained participants were more likely to respond on the basis of *color* than the untrained participants. We ran simulations with AARM using two sets of training stimuli. Here, we discuss simulation results from Training Set A, which had the same characteristics as the training stimuli described above. Results from simulations that used the hypothetical Training Set B will be discussed in Case Study 4B. We first introduced each unique training stimulus to the between-trial module to generate initialized weights for the within-trial module. Because context co-occurred with the relevant target dimension and did not contain independently relevant information, AARM allocated minimal attention to context. As shown in Figure 11E (Training A), the shape and color of the target received the highest attention weights (0.312), and context received the lowest weight (0.040), with shape and color of the choice options falling in between (0.084). After initializing the within-trial model's attention weights, probe stimuli e_i containing each of the two possible context feature values were introduced to the model separately. Within each of the two probes, one choice option matched an arbitrarily chosen target according to shape, and the other choice option matched the target according to color as follows (Table 4).

Using each probe, we ran 1,000 independent within-trial simulations without feedback. Each simulated trial yielded a binary

Figure 11
Case Study 4: Incidental Context



Note. AARM = adaptive attention representation model. Figure adapted with permission from Child Development. (A) Illustration of contexts. (B) Training stimuli. Stimuli were triads of items, and the task was to select one of the two choice options (top two items) that matched the target (bottom item) according to a rule. Yellow arrows indicate a path from the target to the correct choice option. (C) Test stimuli. Note that identical stimulus configurations were shown in Contexts 1 and 2, but yellow arrows indicate that different responses are appropriate according to the context. (D) Hypothetical training stimuli for simulation purposes. (E) Attention weights are generated by the between-trial module of AARM before (left bar) and after exposure to each set of training stimuli and their category labels. Each color represents a stimulus dimension, and larger segment heights correspond to larger attention weights. (F) Observed proportions of shape-based responses in each context. (G and H) Model-generated proportions of shape-based responses in each context at test, following Training A and B. The context dimension in our simulations was either considered to be integrated (perceptually overlapping) with the dimensions of the stimulus triad (G) or segregated (separate in space and requiring independent perceptual processing) from the dimensions of the stimulus triad (H). See the online article for the color version of this figure.

response corresponding to either the left or the right choice option in the stimulus triad.

We assumed that shape and color of a given triad item could be encoded simultaneously, given that they occupied the same location in space. Mechanisms for contextual encoding, however, were much less straightforward. We therefore, performed two sets of simulations using Training Set A, each representing a different hypothesis for how context is processed and encoded. In one set, contextual

Table 4
Probes for Case Study 4

Item	Dimension	Probe 1 features	Probe 2 features
	Context	0	1
Target	Shape	0	0
	Color	0	0
Option 1	Shape	0	0
	Color	1	1
Option 2	Shape	1	1
	Color	0	0

information was considered to be integrated with the item-level information at each respective spatial location, such that the probability of encoding context was updated continuously within trial. In the other set, the context was considered to be a segregated dimension, such that the observer had to fixate to context independently from the items in the triad in order to encode its information. As shown in Panel G, AARM predicts behavioral results that are consistent with the observed data when contextual information is integrated with stimulus information. Specifically, AARM predicts a higher proportion of shape-based responses to test items presented in Context 1 compared to Context 2. When the context is considered to be a segregated dimension, however, Figure 11H shows that AARM does not predict the observed response bias on ambiguous test items following Training A. Instead, the model predicts an approximately equal probability of making a shape-based response in both Contexts 1 and 2. Because context is not independently relevant to the task during training and the observer has therefore not learned to explicitly attend to it, context will only be used during the decision process if it is encoded by other, passive means. Our results suggest that attention in AARM is a possible mechanism for the effects of global context on behavior, such that contextual information can be passively encoded along with the features of a stimulus despite a lack of known predictive utility at the time of learning.

Case Study 4B: Context Relevance

We ran an additional set of simulations using a hypothetical Training Set B, in which context was a hierarchically superordinate indicator of whether shape or color was relevant on each trial. Training B stimuli were configured identically to the test set in Sloutsky and Fisher (2008), such that one choice option matched the target according to shape, and the other choice option matched the target according to color (Figure 11D). Although observed responses using the alternative Training B were not published, we performed these additional simulations to provide direct contrast between the influences of global and local context in AARM's specification. We first used the between-trial module to calculate a set of attention weights after observing all 16 unique stimuli in Training B. As shown in Figure 11E, the context dimension was assigned the highest weight (0.471) followed by target shape and color (0.168) and shape and color of the two choice options (0.048). The posttraining weights from the between-trial module were used to initialize the within-trial model on 1,000 simulations. We used the same two probes that were used to examine the learning effects of Training A.

As previously described, context was implemented as an integrated (passively encoded along with fixated stimulus information) or segregated (encoding requires independent fixation) dimension in two separate sets of simulations. AARM predicts the same pattern of responses at test as a result of Training B, regardless of whether context is considered to be an integrated (Figure 11G) or a segregated (Figure 11H) dimension: In both cases, AARM predicts a higher proportion of shape-based responses when triads are presented in Context 1 compared to Context 2. Because the model is able to learn the hierarchical structure of the Training B stimulus set in which context is the superordinate dimension, it responds to test stimuli by orienting to and encoding context independently from the items in the triad. Therefore, Training B does not require AARM to overcome reduced attention to context via passive encoding related to feature integration.

As part of a study that investigated the context-mediated transfer of learning in categorization tasks, George and Kruschke (2012) used model simulations to demonstrate that the results from Sloutsky and Fisher (2008) could be explained by associative learning alone, without the involvement of additional selective attention mechanisms. More specifically, the authors used two associative learning models (Pearce, 1994; Rescorla & Wagner, 1972) to show that context-consistent responses at test could arise on the basis of asymmetrical feature-level similarity between the given test stimulus and a subset of training items. As shown by our AARM simulations in which context is instantiated as an independent dimension relative to the elements of the stimulus triad, however, the influence of context on behavior is not guaranteed from the experimental design of Sloutsky and Fisher (2008; Figure 11H). If we can, for the purposes of argument, assume that the role of attention is ubiquitous in category learning, AARM's within-trial mechanisms offer an alternative to the purely association-based explanation provided by George and Kruschke (2012) that overcomes reduced attention to context incurred as a result of training. By incorporating global context as an integrated dimension that is peripherally attended during item-level processing at test, AARM predicts context-mediated patterns of behavior consistent with the results of Sloutsky and Fisher (2008; Figure 11G). Although several studies have found evidence that global context during learning influences future decision-making behavior (Geiselman & Glenny, 1977; George & Kruschke, 2012; Murnane & Phelps, 1993, 1994; S. Smith, 1986; S. Smith & Vela, 1992), other studies observed the opposite pattern of results (Griffiths & Le Pelley, 2009; S. Smith & Vela, 2001). Given that AARM makes dissociable predictions about the influence of global context depending on the extent of feature integration, AARM can potentially be used in future work to identify which elements of context are bound to stimuli during encoding, and which are not.

General Discussion

The between-trial module of AARM comprises a theoretical framework for how attention allocation, decision-making, and item representations interact to facilitate learning. Here, we extended AARM to account for within-trial dynamics as well: specifically, the mutually influential timecourses of dimension-level information sampling and response evidence. Like AARM's between-trial module, several models predict learning as a consequence of the strategic manipulation of attentional resources over

the course of a task. Most, however, do not make explicit assumptions about how the latent distribution of attention might affect how dimensions are prioritized within a trial. AARM therefore stands apart from other accounts of category learning because it seeks to close the loop between updating latent attention according to trial-level feedback, and subsequently deploying attentional resources to acquire relevant information when the next stimulus appears.

As discussed in the introductory sections, there were four overarching theoretical components to the current work. First, both within- and between-trial dynamics are described by a common set of mechanisms. For AARM to be viable, it was important that the within-trial module be constructed from the same cognitive machinery and operations that were purported by our previous work to be engaged in service of the broader learning problem (Galdo et al., 2021). As such, the Model Specification section provides a core set of mechanisms that operates at multiple timescales to explain how humans both learn about new categories and acquire information about new stimuli.

Second, humans form simplified representations of stimuli from the features that are perceived to be relevant to the task. Given our previous findings that dimensions compete for attention and that only the attended subset appears to contribute to categorization decisions, the within-trial module was necessary to explain how strategic reorientation and self-termination behaviors might emerge. We therefore specified dynamic processes through which attention, decision evidence, and an evolving stimulus representation inform one another, but only until the observer has acquired enough information about the stimulus to map it to a particular category.

Third, attention allocation is optimized with respect to a goal. The learning problem in the between-trial case is well defined, such that the observer can conceivably redistribute attention upon observation of feedback in an effort to reduce the probability of future errors. Indeed, rational theories of psychological processes predict behaviors via optimization of a cost function given some set of environmental constraints (Sakamoto et al., 2008; Sanborn et al., 2010). The costs of sampling information from a feature that provides support for an incorrect category label cannot be ascertained and avoided, however, before the correct label has been provided by feedback. The within-trial module therefore assumes observers seek additional support for the category label that they believe to be correct at each moment in time. The result is a parsimonious extension to attention optimization that is consistent with observable human biases of confirmatory search.

Fourth, attention processes are sensitive to hierarchical structures. Given eye-tracking results showing distinct temporal ordering effects that are consistent with hierarchical structures (Blair et al., 2009), it was important that the within-trial module be able to produce similar trajectories of orienting. In several of our case studies (i.e., 2, 3, and 4), hierarchical organization of information via selective attention was essential for producing the expected patterns of information sampling behaviors and responses. We argue that hierarchical structures are not a special case of experimental manipulations, but are rather ubiquitous in nature given observable impacts of environmental context on information processing and behavior.

Across four case studies, we used model simulations to demonstrate AARM's capacity for predicting plausible patterns of behavioral responses (Case Studies 3 and 4), eye-tracking data (Case Study 1B), or both simultaneously (Case Studies 1A and 2). Our preliminary results provide qualitative support for the within-trial mechanisms proposed by AARM. In Case Study 1, we demonstrated how individual differences in information sampling and response probabilities could emerge due to selective attention and encoding variability, despite all participants experiencing the same stimuli during training. In Case Study 2, we showed that distinct temporal ordering effects of information sampling emerge in the presence of hierarchical stimuli through a combination of experience-biased orienting and mechanisms for ongoing feature predictions. Case Study 3 used hypothetical stimuli to present the possibility that even when dimensions co-occur in space, selective attention could be a mechanism through which only the information that is relevant to individual trials will be encoded and concurrently contribute to the choice. Case Study 4 explored how contextual features could bias decisions at test even if they were not explicitly attended during training. In the sections to follow, we will discuss the implications of our results and suggest future extensions that pertain to AARM's component mechanisms.

Self-Termination

Most models of category learning assume that observers access all feature information across stimulus dimensions when making category judgments. While this may be plausible in laboratory tasks that include stimuli with only a few dimensions, it is potentially unreasonable to assume that humans encode all available perceptual information from the complex stimuli that they encounter in the real world. To make efficient decisions, humans therefore need to identify the dimensions of information that are relevant to their current goals. Using variants of AARM that instantiated different modes of simplicity bias, Galdo et al. (2021) provided evidence that humans tend toward low-dimensional representations as they learn. One interpretation of these findings is that while humans strive to achieve high accuracy in a task setting, they concurrently seek to reduce time and resource expenditure on individual trials (Boureau et al., 2015; Cisek et al., 2009; Thura et al., 2012; Yau et al., 2021).

Given evidence that memories for past events influence how we make predictions about the environment (S. Smith & Vela, 1992) and encode new information (Bowman & Zeithamova, 2020) it stands to reason that the construction and storage of low-dimensional representations might bear a meaningful impact on how the observer interfaces with new stimuli. In Case Study 1, we used the within-trial module of AARM to investigate the potential impact of feature-level encoding variability on subsequent information sampling behaviors in a paradigm with multiple independently relevant sources of information (Blanco & Sloutsky, 2019). In particular, we manipulated the extent to which previously presented features of the deterministic dimension were successfully encoded in memory, such that they were accessible when the observer forms expectations about what features a new stimulus might take on. If humans form simplified representations based on only a few dimensions, selective attention to a subset of probabilistic dimensions should reduce encoding of deterministic features across trials. Although attention was initialized with the same values across simulations, we found that manipulating feature expectations via the encoding structure of the model was sufficient for predicting

notable differences in fixation paths when new stimuli were presented. Importantly, this manipulation also produced differences in proportions of responses in the presence of novel stimuli with conflicting feature-to-category mappings (Case Study 1A), and the extent of reorientation after a categorization rule change (Case Study 1B) that were consistent with observed effects.

Two contributions of the within-trial module, then, are that it (a) provides an explanation for how low-dimensional representations are formed through self-terminating attention and decision processes and (b) allows us to investigate potential impacts of low-dimensional representations on how observers seek out information and respond when presented with new stimuli. While our interest in the current article was to articulate a theory for how learned information (i.e., memories and goal-directed attention) fundamentally shapes how future knowledge is sought after and acquired, the effects of partial or variable encoding of individual stimuli *during* learning require further investigation. The between-trial module of AARM and other iterations of GCM allow for variable memory strength at the level of the global stimulus, such that traces of exemplars are subject to decay as they recede into the past. It is generally assumed, however, that all features are encoded and are available for similarity comparisons as new stimuli are presented. As a future direction, we will therefore use insights provided by the current work to extend the between-trial module of AARM to problems of partial encoding. In high-dimensional environments in particular, the sources of information that are fixated and encoded early in learning may have profound impacts on how attention is selectively distributed in the future. As such, accounting for partial encoding during the learning process would be essential for assessing the relative contributions of initialized attention weights and feature-level memory in generating patterns of behavior like those observed in Case Study 1.

Confirmatory Search

We have made efforts to contrast AARM with SEA, an alternative theory of learning and information sampling (Braunlich & Love, 2021). As a rational account, SEA's purpose is to identify the most cost-effective action within a set of environmental constraints (Sakamoto et al., 2008). While the two models often make similar predictions, AARM fulfills a different purpose of characterizing plausible mechanisms that manifest in human-like behaviors. Its base implementation was therefore, designed to be amenable to influences from observable biases in human learning, whereas SEA was developed to generate optimal sampling paths under various environmental conditions. One major way that AARM departs from SEA is the specification of confirmatory information search. Although unbiased approaches are demonstrably effective at producing optimal sampling trajectories, behavioral effects of confirmatory search have been widely observed in causal judgment tasks (Rabin & Schrag, 1999; Schustack & Sternberg, 1981; Shaklee & Fischhoff, 1982; Wason & Johnson-Laird, 1972), and more recently in visual search as well (Rajsic et al., 2017; Rajsic et al., 2015).

Although the two models have not been directly compared, both AARM and SEA have been shown to produce human-like behaviors of reorientation and self-termination in the presence of hierarchical stimuli from Blair et al. (2009). The manner in which the models perform the task after training, however, differ in interesting ways. As discussed in the case studies that pertain to hierarchical category structures, AARM's between-trial module upweights attention to the superordinate dimension over the course of training. The within-trial module then orients to the superordinate dimension on the basis of posttraining attention weights. When sufficient cumulative attention is applied for a feature value to be encoded, active retrieval of similar exemplars coupled with ongoing updates to attention causes the observer to reorient to a subordinate dimension, depending on the feature identity that was encoded from the superordinate dimension. After accumulating sufficient evidence for a single category label, the model self-terminates with a response.

Braunlich and Love (2021) performed two sets of simulations of the paradigm from Blair et al. (2009): one using the standard model with full preposterior search and the other using the myopic version of the model. The standard model forecasts all possible sequences of feature values across dimensions at trial onset, calculates the probability of observing each response via cluster activation, and condenses that information into an expected utility of sampling each dimension. The observer then samples information from the dimension with the maximum utility, or terminates the search process in a response if no available dimensions are expected to provide gain beyond a prespecified cost of sampling. The myopic version of the model works similarly to the standard version of SEA, except that feature predictions are made only one step into the future. Given the massive computational load of full preposterior search, the myopic variant of the model was presented in an effort to account for human-like limitations on memory and attention resources.

As shown in Figure 9, AARM predicted trajectories of fixations and dwell times that were consistent with the hierarchical structure of the task, varied appropriately between trial types, and consistently self-terminated after the two trial-relevant features were encoded. We consider the level of detail at which AARM is able to predict behavior to be an advantage of its mechanistic approach; as shown in Figure 9, its predictions closely match the observed timecourse of sampling behavior across participants. SEA, by contrast, only makes predictions about the order in which features are sampled before a response is made. This level of specificity is of course sufficient for a rational account, as the model was designed to determine the probability of discrete actions (e.g., sample a dimension; make a response) given a particular goal and task environment. As such, the standard version of SEA was reported to make predictions that were consistent with observed postlearning behavior insofar as it sampled the superordinate dimension first, self-terminated after sampling the two relevant dimensions on each trial, and correctly categorized items on 93.3% of trials (Braunlich & Love, 2021).

The more parsimonious myopic variant, however, was less successful. Because a one-step forecast produces equal utility predictions across dimensions, the myopic model only oriented to the superordinate dimension on one-third of the trials. Across a majority of trials, the myopic model generated fixation trajectories that were not consistent with the observed effects shown in Figure 9. This discrepancy potentially highlights an important instance in which human behavior departs from the optimal action sequence, even when capacity limitations are considered. The myopic model does not predict effects of initial orientation that are

consistent with the hierarchical design of the task because its balanced, single-step prediction determines that all dimensions are equally likely to support a correct response. While this explanatory issue can be overcome by exhaustive preposterior search, AARM produces the target pattern of behavior by incorporating human-like biases and a nonstationary working representation of the stimulus.

Confirmatory search mechanisms in AARM supported behaviors in Blair et al.'s paradigm that were consistent with rational predictions provided by full preposterior search in SEA, but this approach has potential limitations. For instance, it is often the case that false negatives incur a greater cost than false positives, such that disconfirmatory search would be advantageous. Real-world medical diagnosis is an extreme example, but this balance of costs is relevant to various recognition-primed decision-making tasks as well (Fadde, 2009). Additionally, work investigating search strategies has shown that while people tend to maximize probability gain (i.e., sample dimensions that yield the highest probability of a correct response), strategies that maximize information gain or impact are used in some cases as well (Nelson et al., 2010). Although confirmatory search was an effective way of extending the error-minimization updating rule from the between-trial module of AARM to account for the unsupervised aspect of within-trial dynamics, it may not be a viable solution in all contexts.

Given the diverging theoretical bases of AARM and SEA, a direction of future work will be to conduct quantified comparisons between their predictions. Because SEA determines optimal behaviors with respect to the environment while AARM is more flexible with regard to the influences of individual biases, comparing the predictions of these two models may provide important insight into when and why humans deviate from optimal modes of behavior. One potential avenue is to compare AARM and SEA's predictions in a task like the one designed by Blanco and Sloutsky (2019) and discussed in Case Study 1. Both models can purportedly produce learning traps such that an initially irrelevant dimension continues to be ignored even if it becomes relevant at some point in the future. Nevertheless, the switch from Phase 1 to Phase 2 in the Blanco and Sloutsky (2019) paradigm might provide interesting contrast between AARM and SEA because the optimal behavior is not well defined. When the deterministic dimension is no longer relevant, is it more advantageous to exploit a probabilistic dimension and at least be correct on a subset of trials, or reexplore in order to find the new deterministic dimension?

Endogenous Covert Attention

The proposed AARM framework specifies how latent attention dynamics might give rise to patterns of gaze fixations. It is, therefore, relevant to highlight the theoretical distinction between overt and covert attention as it exists in the visual search literature. Whereas covert attention is a latent psychological construct that may be distributed according to feature salience (exogenous) or in a goal-directed (endogenous) manner, overt attention refers specifically to the movements of the eyes (see Itti & Koch, 2001, for review). Previous work has indicated that overt shifts of attention, or saccades, are preceded by covert shifts in attention resulting from anticipation of a visual target's spatial location (Deubel & Schneider, 1996; Hoffman & Subramaniam, 1995). To explain these results, the influential premotor theory suggests that overt and covert attention are tightly coupled, such that they involve a common set of processing and planning streams and the only difference is that motor processes are specific to overt attention (Rizzolatti et al., 1987, 1994). With these insights in mind, we assumed that latent attention in the within-trial module was continuously updated for all dimensions simultaneously, but fixations were directed to the spatial location corresponding to the most informative dimension. Synchronous updates to latent attention across dimensions, therefore, could result in changes to the fixated location. In light of work demonstrating more successful encoding of task-relevant features that incur selective attention over the course of learning (Deng & Sloutsky, 2015), we additionally specified that feature encoding occurs as a function of cumulative latent attention. With this specification, it is possible to overtly attend to a feature, but to fail to encode it if endogenous covert attention is low.

The decoupling of overt and endogenous covert attention is exemplified by Case Study 3, in which multiple stimulus dimensions could occupy the same location in space, but differed in terms of their relevance to the current trial. In the example, angle, frequency, and context dimensions all overlapped in space and thus could be fixated simultaneously. Nevertheless, as shown in Figure 10B, the angle and frequency dimensions were only encoded when they were necessary for identifying the appropriate category label within the relevant task context. Behavioral and neuroimaging work has supported the idea that humans can selectively attend to a subset of dimensions occupying a common spatial location as well. For example, Rutman et al. (2010) collected electroencephalography (EEG) data while participants viewed overlapping face and scene stimuli. The authors identified differences in event replated potentials (ERPs) that depended on whether participants were cued to focus on the face or the scene, and these differences correlated with subsequent memory for cued and uncued stimulus components (see Gazzaley & Nobre, 2012, for additional review).

Although the current specification of AARM's within-trial module assumed feature encoding was determined from endogenous covert attention alone, influences of exogenous covert attention (driven by bottom-up perceptual salience) are likely to play a role as well. Dugue et al. (2020), for example, recently found that both endogenous and exogenous covert attention facilitate encoding, but endogenous attention uniquely facilitates the read out of feature information. Future work will therefore investigate the extent to which overt attention and feature encoding in AARM should be determined from covert attention in general (i.e., both endogenous and exogenous), or endogenous covert attention specifically. One potential avenue is to contrast fixations to salient features early and late in learning. Studies have shown that overt attention initially orients to salient features, but that these effects can be overcome by increasing endogenous covert attention to task-relevant dimensions (Theeuwes, 2010; Vanunu et al., 2021). With AARM's specification for unconstrained total attention (see the Attention Is Not a Zero-Sum Game section), it would be possible to specify a different baseline attention value for each dimension. This would bias information sampling to salient dimensions early in the task, and overcoming this bias would depend on the observer's ability to explore the other dimensions rather than exploiting information from the salient dimensions alone.

Conclusions

With the specification of the between- and within-trial modules of AARM that were outlined here, we have provided a comprehensive theory for how learning impacts how humans interact with their environment, both in terms of the dimensions they attend and the decisions that they make.

AARM stands apart from previous models of category learning in that it presents a common set of mechanisms that operate at both between- and within-trial timescales of attention allocation and decision-making. Our theory broadly suggests that as humans learn, they make decisions on the basis of simplified representations of the stimuli they encounter. These simplified representations gradually emerge through a combination of selective attention to relevant dimensions, and early termination of information search when an evidence threshold is reached. Accumulation of category evidence occurs concurrently with confirmatory information search, such that humans intuitively direct their attention toward dimensions that are expected to support their current beliefs. When testing AARM's theoretical predictions, we focused on hierarchical category structures in particular, due to the natural emergence of temporal ordering effects alongside attention updating. Beyond the results presented here, we believe that AARM comprises a broader theoretical statement about how humans learn in naturalistic environments as well, with contextual dimensions serving as superordinate cues to guide information sampling. This work therefore serves to highlight aspects of category learning that are frequently overlooked, but are crucial for gaining a complete understanding of how humans acquire knowledge about the world.

References

Addleman, D., Tao, J., Remington, R., & Jiang, Y. (2018). Explicit goal-driven attention, unlike implicitly learned attention, spreads to secondary tasks. *Journal of Experimental Psychology: Human Perception and Performance*, *44*(3), Article 356. https://doi.org/10.1037/xhp0000457

Allport, A., Styles, E., & Hsieh, S. (1994). Shifting intentional set: Exploring the dynamic control of tasks. In C. Umilta & M. Moscovitch (Eds.), *Attention and performance XV: Conscious and nonconscious information processing* (pp. 421–452). MIT Press.

Anderson, B., Laurent, P., & Yantis, S. (2011). Value-driven attentional capture. *Proceedings of the National Academy of Sciences*, *108*(25), 10367–10371. https://doi.org/10.1073/pnas.1104047108

Anderson, J. (1991a). The adaptive nature of human categorization. *Psychological Review*, *98*(3), 409–429. https://doi.org/10.1037/0033-295X.98.3.409

Anderson, J. (1991b). Is human cognition adaptive? *Behavioral & Brain Sciences*, *14*(3), 471–484. https://doi.org/10.1017/S0140525X00070801

Ashby, G., Alfonso-Reese, L., & Waldron, E. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review*, *105*(3), 442–481. https://doi.org/10.1037/0033-295X.105.3.442

Ashby, G., & Maddox, T. (2005). Human category learning. *Annual Review of Psychology*, *56*(1), 149–178. https://doi.org/10.1146/psych.2005.56.issue-1

Awh, E., Vogel, E., & Oh, S.-H. (2006). Interactions between attention and working memory. *Neuroscience*, *139*(1), 201–208. https://doi.org/10.1016/j.neuroscience.2005.08.023

Baddeley, A. (1982). Domains of recollection. *Psychological Review*, *89*(6), 708–729. https://doi.org/10.1037/0033-295X.89.6.708

Baker, A., Kim, M., & Hoffman, J. (2021). Searching for emotional salience. *Cognition*, *214*, Article 104730. https://doi.org/10.1016/j.cognition.2021.104730

Barto, A., & Mahadevan, S. (2003). Recent advances in hierarchical reinforcement learning. *Discrete Event Dynamic Systems*, *13*(1), 41–77. https://doi.org/10.1023/A:1022140919877

Blair, M., Watson, M., Walshe, R., & Maj, F. (2009). Extremely selective attention: Eye-tracking studies of the dynamic allocation of attention to stimulus features in categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(5), 1196–1206. https://doi.org/10.1037/a0016272

Blanco, N., & Sloutsky, V. (2019). Adaptive flexibility in category learning? Young children exhibit smaller costs of selective attention than adults. *Developmental Psychology*, *55*(10), 2060–2076. https://doi.org/10.1037/dev0000777

Blanco, N., Turner, B., & Sloutsky, V. (under review). *The benefits of immature cognitive control: How distrubuted attention guards against learning traps*.

Bonardi, C., Graham, S., Hall, G., & Mitchell, C. (2005). Acquired distinctiveness and equivalence in human discrimination learning: Evidence for an attentional process. *Psychonomic Bulletin & Review*, *12*(1), 88–92. https://doi.org/10.3758/BF03196351

Botta, F., Martin-Arevalo, E., Lupianez, J., & Bartolomeo, P. (2019). Does spatial attention modulate sensory memory? *PLOS ONE*, *14*(7), Article e0219504. https://doi.org/10.1371/journal.pone.0219504

Botvinick, M. (2012). Hierarchical reinforcement learning and decision making. *Current Opinion in Neurobiology*, *22*(6), 956–962. https://doi.org/10.1016/j.conb.2012.05.008

Botvinick, M., Niv, Y., & Barto, A. (2009). Hierarchically organized behavior and its neural foundations: A reinforcement learning perspective. *Cognition*, *113*(3), 262–280. https://doi.org/10.1016/j.cognition.2008.08.011

Boureau, Y., Sokol-Hessner, P., & Daw, N. (2015). Deciding how to decide: Self-conrol and meta-decision making. *Trends in Cognitive Sciences*, *19*(11), 700–710. https://doi.org/10.1016/j.tics.2015.08.013

Bowman, C., & Zeithamova, D. (2018). Abstract memory representations in the ventromedial prefrontal cortex and hippocampus support concept generalization. *Journal of Neuroscience*, *38*(10), 2605–2614. https://doi.org/10.1523/JNEUROSCI.2811-17.2018

Bowman, C., & Zeithamova, D. (2020). Training set coherence and set size effects on concept generalization and recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(8), 1442–1464. https://doi.org/10.1037/xlm0000824

Braunlich, K., & Love, B. (2021). Bidirectional influences of information-sampling and concept learning. *Psychological Review*, *129*(2), 213–234. https://doi.org/10.1037/rev0000287

Brockdorff, N., & Lamberts, K. (2000). A feature-sampling account of the time course of old-new recognition judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(1), 77–102. https://doi.org/10.1037/0278-7393.26.1.77

Brown, S., & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology*, *57*(3), 153–178. https://doi.org/10.1016/j.cogpsych.2007.12.002

Brydges, C., Clunies-Ross, K., Clohessy, M., Lo, Z., Nguyen, A., Rousset, C., Whitelaw, P., Yeap, Y. J., & Fox, A. M. (2012). Dissociable components of cognitive control: An event-related potential (erp) study of response inhibition and interference suppression. *PLOS ONE*, *7*(3), Article e34428. https://doi.org/10.1371/journal.pone.0034482

Busemeyer, J., & Townsend, J. (1993). Decision field theory: A dynamic-cognitive approach to decision making in an uncertain environment. *Psychological Review*, *100*(3), Article 492. https://doi.org/10.1037/0033-295X.100.3.432

Chen, L., Meier, K., Blair, M., Watson, M., & Wood, M. (2013). Temporal characteristics of overt attentional behavior during category learning. *Attention, Perception, & Psychophysics*, *75*(2), 244–256. https://doi.org/10.3758/s13414-012-0395-8

Chiu, Y.-C., & Yantis, S. (2009). A domain-independent source of cognitive control for task sets: Shifting spatial attention and switching categorization rules. *Journal of Neuroscience*, *29*(12), 3930–3938. https://doi.org/10.1523/JNEUROSCI.5737-08.2009

Chun, M., Golomb, J., & Turk-Browne, N. (2011). A taxonomy of external and internal attention. *Annual Review of Psychology*, *62*(1), 73–101. https://doi.org/10.1146/annurev.psych.093008.100427

Chun, M., & Jiang, Y. (1998). Contextual cueing: Implicit learning and memory of visual context guides spatial attention. *Cognitive Psychology*, *36*(1), 28–71. https://doi.org/10.1006/cogp.1998.0681

Chun, M., & Turk-Browne, N. (2007). Interactions between attention and memory. *Current Opinion in Neurobiology*, *17*(2), 177–184. https://doi.org/10.1016/j.conb.2007.03.005

Cisek, P., Puskas, G., & El-Lurr, S. (2009). Decisions in changing conditions: The urgency-gating model. *Journal of Neuroscience*, *29*(37), 11560–11571. https://doi.org/10.1523/JNEUROSCI.1844-09.2009

Cohen, A., & Nosofsky, R. (2003). An extension of the exemplar-based random-walk model to separable-dimension stimuli. *Journal of Mathematical Psychology*, *47*(2), 150–165. https://doi.org/10.1016/S0022-2496(02)00031-7

Cohen, D., Dunbar, K., & McClelland, J. (1990). On the control of automatic processes: A parallel distributed processing account of the stroop effect. *Psychological Review*, *97*(3), Article 332. https://doi.org/10.1037/0033-295X.97.3.332

Cox, G., & Criss, A. (2020). Similarity leads to correlated processing: A dynamic model of encoding and recognition of episodic associations. *Psychological Review*, *127*(5), 792–828. https://doi.org/10.1037/rev0000195

Cox, G., & Shiffrin, R. (2017). A dynamic approach to recognition memory. *Psychological Review*, *124*(6), Article 795. https://doi.org/10.1037/rev0000076

Crump, M., Milliken, B., Leboe-McGowan, J., Lebowe-McGowan, L., & Gao, X. (2018). Context-dependent control of attention capture: Evidence from proportion congruent effects. *Canadian Journal of Experimental Psychology*, *72*(2), 91–104. https://doi.org/10.1037/cep0000145

De Brigard, F., Addis, D., Ford, J., Schacter, D., & Giovanello, K. (2013). Remembering what could have happened: Neural correlates of episodic counterfactual thinking. *Neuropsychologia*, *51*(12), 2401–2414. https://doi.org/10.1016/j.neuropsychologia.2013.01.015

De Brigard, F., Giovanello, K., Stewart, G., Lockrow, A., O'Brien, M., & Spreng, R. (2016). Characterizing the subjective experience of episodic past, future, and counterfactual thinking in healthy younger and older adults. *Quarterly Journal of Experimental Psychology*, *69*(12), 2358–2375. https://doi.org/10.1080/17470218.2015.1115529

De Brigard, F., Spreng, R., Mitchell, J., & Schacter, D. (2015). Neural activity associated with self, other, and object-based counterfactual thinking. *NeuroImage*, *109*(2), 12–26. https://doi.org/10.1016/j.neuroimage.2014.12.075

Deese, J. (1959). On the prediction of occurrence of particular verbal intrusions in immediate recall. *Journal of Experimental Psychology*, *58*(1), 17–22. https://doi.org/10.1037/h0046671

Demirkaya, A., Chen, J., & Symak, S. (2020, March). *Exploring the role of loss functions in multiclass classification* [Conference session]. In 2020 54th annual conference on information sciences and systems (ciss) IEEE.

Deng, W., & Sloutsky, V. (2015). The development of categorization: Effects ofclassification and inference training on category representation. *Developmental Psychology*, *51*(3), 392–405. https://doi.org/10.1037/a0038749

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, *18*(1), 193–222. https://doi.org/10.1146/annurev.ne.18.030195.001205

Deubel, H., & Schneider, W. (1996). Saccade target selection and object recognition: Evidence for a common attentional mechanism. *Vision Research*, *36*(12), 1827–1837. https://doi.org/10.1016/0042-6989(95)00294-4

Dix, S., & Aggleton, J. (1999). Extending the spontaneous preference test of recognition: Evidence of object-location and object-context recognition. *Behavioral Brain Research*, *99*(2), 191–200. https://doi.org/10.1016/S0166-4328(98)00079-5

Dosher, B. (1984). Discriminating preexperimental (semantic) from learned (episodic) associations: A speed-accuracy study. *Cognitive Psychology*, *16*(4), 519–555. https://doi.org/10.1016/0010-0285(84)90019-7

Dosher, B., & Rosedale, G. (1991). Judgments of semantic and episodic relatedness: Common time-course and failure of segregation. *Journal of Memory and Language*, *30*(2), 125–160. https://doi.org/10.1016/0749-596X(91)90001-Z

Doucet, A., de Freitas, N., & Gordon, N. (2001). *Sequential Monte Carlo methods in practice*. Springer.

Dugue, L., Merriam, E., Heeger, D., & Carrasco, M. (2020). Differential impact of endogenous and exogenous attention on activity in human visual cortex. *Scientific Reports*, *10*(21274), Article 1484. https://doi.org/10.1038/s41598-020-78172-x

Dulsky, S. (1935). The effect of a change of background on recall and relearning. *Journal of Experimental Psychology*, *18*(6), 725–740. https://doi.org/10.1037/h0058066

Egner, T. (2008). Multiple conflict-driven control mechanisms in the human brain. *Trends in Cognitive Sciences*, *12*(10), 374–380. https://doi.org/10.1016/j.tics.2008.07.001

Eich, E. (1985). Context, memory, and integrated item/context imagery. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*(4), 764–770. https://doi.org/10.1037/0278-7393.11.1-4.764

Estes, W. (1972). An associative basis for coding and organization in memory. In A. Melton & E. Martin (Eds.), *Coding processes in human memory* (pp. 161–190). Winston & Sons.

Estes, W. (1986). Array models for category learning. *Cognitive Psychology*, *18*(4), 500–549. https://doi.org/10.1016/0010-0285(86)90008-3

Estes, W. (1994). *Classification and cognition*. Oxford University Press.

Fadde, P. (2009). Instructional design for advanced learners: Training recognition skills to hasten expertise. *Educational Technology, Research and Development*, *57*(3), 359–376. https://doi.org/10.1007/s11423-007-9046-5

Foster, J., Bsales, E., & Awh, E. (2020). Covert spatial attention speeds target individuation. *Journal of Neuroscience*, *40*(13), 2717–2726. https://doi.org/10.1523/JNEUROSCI.2962-19.2020

Galdo, M., Weichart, E., Sloutsky, V., & Turner, B. (2021). *The quest for simplicity in human learning: Identifying the constraints on attention*. https://doi.org/10.31234/osf.io/xgfmb

Gazzaley, A., & Nobre, A. (2012). Top-down modulation: Bridging selective attention and working memory. *Trends in Cognitive Sciences*, *16*(2), 129–135. https://doi.org/10.1016/j.tics.2011.11.014

Geiselman, R., & Glenny, J. (1977). Effects of imagining speakers' voices on the retention of words presented visually. *Memory & Cognition*, *5*(5), 499–504. https://doi.org/10.3758/BF03197392

George, D., & Kruschke, J. (2012). Contextual modulation of attention in human category learning. *Learning & Behavior*, *40*(4), 530–541. https://doi.org/10.3758/s13420-012-0072-8

Gilks, W., Richardson, S., & Spiegelhalter, J. (1996). *Markov chain Monte Carlo in practice*. Chapman and Hall.

Gluck, M., & Bower, G. (1988). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General*, *117*(3), Article 227. https://doi.org/10.1037/0096-3445.117.3.227

Godden, D., & Baddeley, A. (1975). Context-dependent memory in two natural environments: On land and underwater. *British Journal of Psychology*, *66*(3), 325–331. https://doi.org/10.1111/j.2044-8295.1975.tb01468.x

Goldberg, P., & Jerrum, M. (1995). Bounding the VC dimension of concept classes parameterized by real numbers. *Machine Learning*, *18*(2–3), 131–148. https://doi.org/10.1007/BF00993408

Goodfellow, I. (2016). *Deep learning* (Vol. 1, No. 2). MIT Press.

Greene, R., & Tussing, A. (2001). Similarity and associative recognition. *Journal of Memory and Language*, *45*(4), 573–584. https://doi.org/10.1006/jmla.2001.2795

Griffiths, O., & Le Pelley, M. (2009). Attentional changes in blocking are not a consequence of lateral inhibition. *Learning & Behavior*, *37*(1), 27–41. https://doi.org/10.3758/LB.37.1.27

Hall, G. (1991). *Perceptual and associative learning*. Oxford University Press.

Hintzman, D. (1984). Minerva 2: A simulation model of human memory. *Behavior Research Methods, Instruments, & Computers*, *16*(2), 96–101. https://doi.org/10.3758/BF03202365

Hockley, W. (2008). The effects of environmental context on recognition memory and claims of remembering. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(6), 1412–1429. https://doi.org/10.1037/a0013016

Hoffman, A., & Rehder, B. (2010). The costs of supervised classification: The effect of learning task on conceptual flexibility. *Journal of Experimental Psychology:General*, *139*(2), 319–340. https://doi.org/10.1037/a0019042

Hoffman, J., & Subramaniam, B. (1995). The role of visual attention in saccadic eye movements. *Perception & Psychophysics*, *57*(6), 787–795. https://doi.org/10.3758/BF03206794

Hunsaker, M., & Kesner, R. (2013). The operation of pattern separation and pattern completion processes associated with different attributes or domains of memory. *Neuroscience & Biobehavioral Reviews*, *37*(1), 36–58. https://doi.org/10.1016/j.neubiorev.2012.09.014

Irwin, D. (1996). Integrating information across saccadic eye movements. *Current Directions in Psychological Science*, *5*(3), 94–100. https://doi.org/10.1111/1467-8721.ep10772833

Isarida, T., & Isarida, T. (2007). Environmental context effects of background color in free recall. *Memory & Cognition*, *35*(7), 1620–1629. https://doi.org/10.3758/BF03193496

Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews Neuroscience*, *2*(3), 194–203. https://doi.org/10.1038/35058500

Janssens, C., De Loof, E., Boehler, N., Pourtois, G., & Verguts, T. (2018). Occipital alpha power reveals fast attentional inhibition of incongruent distractors. *Psychophysiology*, *55*(3), Article e13011. https://doi.org/10.1111/psyp.13011

Johnston, W., & Schwarting, I. (1997). Novel popout: An enigma for conventional theories of attention. *Journal of Experimental Psychology: Human Perception and Performance*, *23*(3), Article 622. https://doi.org/10.1037/0096-1523.23.3.622

Krajbich, I., Armel, C., & Rangel, A. (2010). Visual fixations and the computation and comparison of value in simple choice. *Nature Neuroscience*, *13*(10), 1292–1298. https://doi.org/10.1038/nn.2635

Krajbich, I., & Rangel, A. (2011). Multialternative drift-diffusion model predicts the relationship between visual fixations and choice in value-based choice. *Proceedings of the National Academy of Sciences of the United States of America*, *108*, 13852–13857. https://doi.org/10.1073/pnas.110132810

Kruschke, J. (1992). Alcove: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*(1), 22–44. https://doi.org/10.1037/0033-295X.99.1.22

Kruschke, J. (1996). Dimensional relevance shifts in category learning. *Connection Science*, *8*(2), 225–248. https://doi.org/10.1080/095400996116893

Kruschke, J. (2001). Toward a unified model of attention in associative learning. *Journal of Mathematical Psychology*, *45*(6), 812–863. https://doi.org/10.1006/jmps.2000.1354

Kumaran, D., & McClelland, J. (2012). Generalization through the recurrent interaction of episodic memories: A model of the hippocampal system. *Psychological Review*, *119*(3), 573–616. https://doi.org/10.1037/a0028681

Lai, M.-L., Tsai, M.-J., Yang, F.-Y., Hsu, C.-Y., Liu, T.-C., Lee, S., Chiou, G. L., Liang, J. C. & Tsai, C.-C. (2013). A review of using eye-tracking technology in exploring learning from 2000 to 2012. *Educational Research Review*, *10*(5), 90–115. https://doi.org/10.1016/j.edurev.2013.10.001

Lamberts, K. (2000). Information-accumulation theory of speeded categorization. *Psychological Review*, *107*(2), 227–260. https://doi.org/10.1037/0033-295X.107.2.227

Lashley, K. (1951). The problem of serial order in behavior. In L. Jeffress (Ed.), *Cerebral mechanisms in behavior: The Hixon symposium* (pp. 112–136). Wiley.

Lavie, N. (1995). Perceptual load as a necessary condition for selective attention. *Journal of Experimental Psychology: Human Perception and Performance*, *21*(4), Article 451. https://doi.org/10.1037/0096-1523.21.3.451

Lavie, N., & Cox, S. (1997). On the efficiency of visual selective attention: Efficient visual search leads to inefficient distractor rejection. *Psychological Science*, *8*(5), 395–396. https://doi.org/10.1111/j.1467-9280.1997.tb00432.x

Lavie, N., & Tsal, Y. (1994). Perceptual load as a major determinant of the locus of selection in visual space. *Perception & Psychophysics*, *56*(2), 183–197. https://doi.org/10.3758/BF03213897

Le Pelley, M. (2004). The role of associative history in models of associative learning: A selective review and hybrid model. *Quarterly Journal of Experimental Psychology*, *57B*, 193–243. https://doi.org/10.1080/02724990344000141

Lee, M. (2001). On the complexity of additive clustering models. *Journal of Mathematical Psychology*, *45*(1), 131–148. https://doi.org/10.1006/jmps.1999.1299.

Lefebvre, G., Summerfield, C., & Bogacz, R. (2022). A normative account of computation bias during reinforcement learning. *Neural Computation*, *34*(2), 307–337. https://doi.org/10.1162/neco_a_01455

Loftus, G. (1985). Picture perception: Effects of luminance on available information and information-extraction rate. *Journal of Experimental Psychology: General*, *114*(3), Article 342. https://doi.org/10.1037/0096-3445.114.3.342

Logan, G. (1988). Toward an instance theory of automatization. *Psychological Review*, *95*(4), Article 492. https://doi.org/10.1037/0033-295X.95.4.492

Logan, G. (2002). An instance theory of attention and memory. *Psychological Review*, *109*(2), Article 376. https://doi.org/10.1037/0033-295X.109.2.376

Love, B., Medin, D., & Gureckis, T. (2004). Sustain: A network model of category learning. *Psychological Review*, *111*(2), Article 309. https://doi.org/10.1037/0033-295X.111.2.309.

Mack, M., Love, B., & Preston, A. (2016). Dynamic updating of hippocampal object representations relfects new conceptual knowledge. *Proceedings of the National Academy of Sciences*, *113*(46), 13203–13208. https://doi.org/10.1073/pnas.1614048113

Mack, M., Love, B., & Preston, A. (2018). Building concepts one episode at a time: The hippocampus and concept formation. *Neuroscience Letters*, *260*(44), 31–38. https://doi.org/10.1016/j.neulet.2017.07.061

Mack, M., Preston, A., & Love, B. (2013). Decoding the brain's algorithm for categorization from its neural implementation. *Current Biology*, *23*(20), 2023–2027. https://doi.org/10.1016/j.cub.2013.08.035

Mackintosh, N. (1975). A theory of attention: Variations in the associability of stimuli with reinforcement. *Psychological Review*, *82*(4), Article 276. https://doi.org/10.1037/h0076778

Mackintosh, N., & Little, L. (1969). Intradimensional and extradimensional shift learning by pigeons. *Psychonomic Science*, *14*(1), 5–6. https://doi.org/10.3758/BF03336395

Maddox, T., & Ashby, G. (2004). Dissociating explicit and procedural-learning based systems of perceptual category learning. *Behavioral Processes*, *66*(3), 309–332. https://doi.org/10.1016/j.beproc.2004.03.011

Markman, A., & Ross, B. (2003). Category use and category learning. *Psychological Bulletin*, *129*(4), 592–613. https://doi.org/10.1037/0033-2909.129.4.592

McColeman, C., Barnes, J., Chen, L., Meier, K., Walshe, R., & Blair, M. (2014). Learning-induced changes in attentional allocation during categorization: A sizable catalog of attention change as measured by eye movements. *PLOS ONE*, *9*(1), Article e83302. https://doi.org/10.1371/journal.pone.0083302

McMillen, T., & Holmes, P. (2006). The dynamics of choice among multiple alternatives. *Journal of Mathematical Psychology*, *50*(1), 30–57. https://doi.org/10.1016/j.jmp.2005.10.003

Medin, D., & Schaffer, M. (1978). Context theory of classification learning. *Psychological Review*, *85*(3), Article 207. https://doi.org/10.1037/0033-295X.85.3.207

Meier, K., & Blair, M. (2013). Waiting and weighting: Information sampling is a balance between efficiency and error-reduction. *Cognition*, *126*(2), 219–325. https://doi.org/10.1016/j.cognition.2012.09.014

Meiran, N. (1996). Reconfiguration of processing mode prior to task performance. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(6), 1423–1442. https://doi.org/10.1037/0278-7393.22.6.1423

Miller, G., Galanter, E., & Pribram, K. (1960). *Plans and the structure of behavior*. Holt, Rinehart, & Winston.

Monsell, S. (2003). Task switching. *Trends in Cognitive Sciences*, *7*(3), 134–140. https://doi.org/10.1016/S1364-6613(03)00028-7

Most, S., Chun, M., Widders, D., & Zald, D. (2005). Attentional rubbernecking: Cognitive control and personality in emotion-induced blindness. *Psychonomic Bulletin & Review*, *12*(4), 654–661. https://doi.org/10.3758/BF03196754

Muller, H., & von Muhlenen, A. (2000). Probing distractor inhibition in visual search: Inhibition of return. *Journal of Experimental Psychology: Human Perception and Performance*, *26*(5), 1591–1605. https://doi.org/10.1037/0096-1523.26.5.1591

Muller, H., von Muhlenen, A., & Geyer, T. (2007). Top-down inhibition of search distractors in parallel visual search. *Perception & Psychophysics*, *69*(8), 1373–1388. https://doi.org/10.3758/BF03192953

Murnane, K., & Phelps, M. (1993). A global activation approach to the effect of changes in environmental context on recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(4), 882–894. https://doi.org/10.1037/0278-7393.19.4.882

Murnane, K., & Phelps, M. (1994). When does a different environmental context make a difference in recognition? a global activation model. *Memory & Cognition*, *22*(5), 584–590. https://doi.org/10.3758/BF03198397

Murnane, K., Phelps, M., & Malmberg, K. (1999). Context-dependent recognition memory: The ice theory. *Journal of Experimental Psychology: General*, *128*(4), 403–415. https://doi.org/10.1037/0096-3445.128.4.403

Nelson, J., & Cottrell, G. (2007). A probabilistic model of eye movements in concept formation. *Neurocomputing*, *70*(13–15), 2256–2272. https://doi.org/10.1016/j.neucom.2006.02.026

Nelson, J., McKenzie, C., Cottrell, G., & Sejnowski, T. (2010). Experience matters: Information acquisition optimizes probability gain. *Psychological Science*, *21*(7), 960–969. https://doi.org/10.1177/0956797610372637

Nickerson, R. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*(2), 175–220. https://doi.org/10.1037/1089-2680.2.2.175

Niwa, M., & Ditterich, J. (2008). Perceptual decisions between multiple directions of visual motion. *Journal of Neuroscience*, *28*(17), Article 4435. https://doi.org/10.1523/JNEUROSCI.5564-07.2008

Norman, D. (1968). Toward a theory of memory and attention. *Psychological Review*, *75*(6), 522–536. https://doi.org/10.1037/h0026699

Norman, D. A., & Shallice, T. (1986). Attention to action. In R. J. Davidson, G. E. Schwartz, & D. Shapiro (Eds.), *Consciousness and self-regulation*. Springer. https://doi.org/10.1007/978-1-4757-0629-1_1.

Nosofsky, N., Little, D., & James, T. (2012). Activation in the neural network responsible for categorization and recognition reflects parameter changes. *Proceedings of the National Academy of Sciences*, *109*(1), 333–338. https://doi.org/10.1073/pnas.1111304109

Nosofsky, R. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39–57. https://doi.org/10.1037/0096-3445.115.1.39

Nosofsky, R. (1988). Exemplar-based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(4), 700–708. https://doi.org/10.1037/0278-7393.14.4.700

Nosofsky, R., & Palmeri, T. (1997). An exemplar-based random walk model of speeded classification. *Psychological Review*, *104*(2), 266–300. https://doi.org/10.1037/0033-295X.104.2.266

Nosofsky, R., Palmeri, T., & McKinley, S. (1994). Rule-plus-exception model of classification learning. *Psychological Review*, *101*(1), 53–79. https://doi.org/10.1037/0033-295X.101.1.53

Oaksford, M., & Chater, N. (Eds.). (1998). *Rational models of cognition*. Oxford University Press.

O'Donoghue, E., Broschard, M., & Wasserman, E. (2020). Pigeons exhibit flexibility but not rule formation in dimensional learning, stimulus generalization, and task switching. *Journal of Experimental Psychology: Animal Learning and Cognition*, *46*(2), 107–123. https://doi.org/10.1037/xan0000234

Olivers, C., Peters, J., Houtkamp, R., & Roelfsema, P. (2011). Different states in visual working memory: When it guides attention and when it does not. *Trends in Cognitive Sciences*, *15*(7), 327–334. https://doi.org/10.1016/j.tics.2011.05.004

Palmeri, T. (1999). Learning categories at different hierarchical levels: A comparison of category learning models. *Psychonomic Bulletin & Review*, *6*(3), 495–503. https://doi.org/10.3758/BF03210840

Park, T., & Casella, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association*, *103*(482), 681–686. https://doi.org/10.1198/016214508000000337

Paskewitz, S., & Jones, M. (2020). Dissecting EXIT. *Journal of Mathematical Psychology*, *97*(25), Article 102371. https://doi.org/10.1016/j.jmp.2020.102371

Pearce, J. (1994). Similarity and discrimination: A selective review and a connectionist model. *Psychological Review*, *101*(4), 587–607. https://doi.org/10.1037/0033-295X.101.4.587

Perfect, T. (1996). Does context discriminate recollection from familiarity in recognition memory. *The Quarterly Journal of Experimental Psychology: Section A*, *49*(3), 797–813. https://doi.org/10.1080/713755644

Pooley, J., Lee, M., & Shankle, W. (2011). Understanding memory impairment with memory models and hierarchical Bayesian analysis. *Journal of Mathematical Psychology*, *55*(1), 47–56. https://doi.org/10.1016/j.jmp.2010.08.003

Rabin, M., & Schrag, J. (1999). First impressions matter: A model of confirmatory bias. *The Quarterly Journal of Economics*, *114*(1), 37–82. https://doi.org/10.1162/003355399555945

Rajsic, J., Taylor, E., & Pratt, J. (2017). Out of sight, out of mind: Matching bias underlies confirmatory visual search. *Attention, Perception, & Psychophysics*, *79*(2), 498–507. https://doi.org/10.3758/s13414-016-1259-4

Rajsic, J., Wilson, D., & Pratt, J. (2015). Confirmation bias in visual search. *Journal of Experimental Psychology: Human Perception and Performance*, *41*(5), 1353–1364. https://doi.org/10.1037/xhp0000090

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59–108. https://doi.org/10.1037/0033-295X.85.2.59

Rehder, B., & Hoffman, A. (2005a). Eyetracking and selective attention in category learning. *Cognitive Psychology*, *51*(1), 1–41. https://doi.org/10.1016/j.cogpsych.2004.11.001

Rehder, B., & Hoffman, A. (2005b). Thirty-something categorization results explained: Attention, eyetracking, and models of category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(5), 811–829. https://doi.org/10.1037/0278-7393.31.5.811

Rescorla, R., & Wagner, A. (1972). A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In

A. Black & W. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (pp. 64–99). Appleton-Century-Crofts.

Rich, A., & Gureckis, T. (2018). Exploratory choice reflects the future value of information. *Decision*, *5*(3), Article 177. https://doi.org/10.1037/dec0000074

Richter, F., Chanales, A., & Kuhl, B. (2016). Predicting the integration of overlapping memories by decoding mnemonic processing states during learning. *NeuroImage*, *124*(3), 323–335. https://doi.org/10.1016/j.neuroimage.2015.08.051

Rizzolatti, G., Riggio, L., Dascola, I., & Umilta, C. (1987). Reorienting attention across the horizontal and vertical meridians: Evidence in favor of a premotor theory of attention. *Neuropsychologia*, *25*(1), 31–40. https://doi.org/10.1016/0028-3932(87)90041-8

Rizzolatti, G., Riggio, L., & Sheliga, B. (1994). Space and selective attention. In C. Umilta & M. Moscovitch (Eds.), *Attention and performance XV: Conscious and nonconscious information processing* (pp. 231–264). MIT Press.

Roediger, H., & McDermott, K. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *21*(4), 803–814. https://doi.org/10.1037/0278-7393.21.4.803

Rogers, R., & Monsell, S. (1995). Costs of a predictable switch between simple cognitive tasks. *Journal of Experimental Psychology: General*, *124*(2), 207–231. https://doi.org/10.1037/0096-3445.124.2.207

Rumelhart, D., & McClelland, J. (1988). *Parallel distributed processing*. IEEE.

Rutman, A., Clapp, W., Chadick, J., & Gazzaley, A. (2010). Early top-down control of visual processing predicts working memory performance. *Journal of Cognitive Neuroscience*, *22*(6), 1224–1234. https://doi.org/10.1162/jocn.2009.21257

Sakamoto, Y., Jones, M., & Love, B. (2008). Putting the psychology back into psychological models: Mechanistic versus rational approaches. *Memory &Cognition*, *36*(6), 1057–1065. https://doi.org/10.3758/MC.36.6.1057

Sanborn, A., Griffiths, T., & Navarro, D. (2010). Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review*, *117*(4), Article 1144. https://doi.org/10.1037/a0020511

Schapiro, A., Turk-Browne, N., Botvinick, M., & Norman, K. (2017). Complementary learning systems within the hippocampus: A neural network modelling approach to reconciling episodic memory with statistical learning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1711), Article 20160049. https://doi.org/10.1098/rstb.2016.0049

Schustack, M., & Sternberg, R. (1981). Evaluation of evidence in causal inference. *Journal of Experimental Psychology: General*, *110*(1), 101–120. https://doi.org/10.1037/0096-3445.110.1.101

Shaklee, H., & Fischhoff, B. (1982). Strategies of information search in causal analysis. *Memory & Cognition*, *10*(6), 520–530. https://doi.org/10.3758/BF03202434

Shepard, R. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*, 1317–1323. https://doi.org/10.1126/science.3629243

Shepard, R., & Arabie, P. (1979). Additive clustering: Representation of similarities as a combination of discrete overlapping properties. *Psychological Review*, *86*(2), 87–123. https://doi.org/10.1037/0033-295X.86.2.87

Shepard, R., Hovland, C., & Jenkins, H. (1961). Learning and memorization of classifications. *Psychological Mongraphs: General and Applied*, *75*(13), 1–42. https://doi.org/10.1037/h0093825

Shiffrin, R., & Schneider, W. (1977). Controlled and automatic human information processing: Ii. perceptual learning, automatic attending and a general theory. *Psychological Review*, *84*(2), 127–190. https://doi.org/10.1037/0033-295X.84.2.127

Shiffrin, R., & Steyvers, M. (1997). A model for recognition memory: Rem–retrieving effectively from memory. *Psychonomic Bulletin & Review*, *4*(2), 145–166. https://doi.org/10.3758/BF03209391

Sloutsky, V. (2003). The role of similarity in the development of categorization. *Trends in Cognitive Science*, *7*(6), 246–558. https://doi.org/10.1016/S1364-6613(03)00109-8

Sloutsky, V., & Fisher, A. (2008). Attentional learning and flexible induction: How mundane mechanisms give rise to smart behaviors. *Child Development*, *79*(3), 639–651. https://doi.org/10.1111/j.1467-8624.2008.01148.x

Smith, D., Berg, M., Cook, R., Murphy, M., Crossley, M., Boomer, J., & Spiering, B. (2012). Implicit and explicit categorization: A tale of four species. *Neuroscience & Biobehavioral Reviews*, *36*(10), 2355–2369. https://doi.org/10.1016/j.neubiorev.2012.09.003

Smith, S. (1986). Environmental context-dependent recognition memory using a short-term memory task for input. *Memory & Cognition*, *14*(4), 347–354. https://doi.org/10.3758/BF03202513

Smith, S., Glenberg, A., & Bjork, R. (1978). Environmental context and human memory. *Memory & Cognition*, *6*(4), 342–353. https://doi.org/10.3758/BF03197465

Smith, S., & Krajbich, I. (2019a). Gaze amplifies value in decision making. *Psychological Science*, *30*(1), 116–128. https://doi.org/10.1177/0956797618810521

Smith, S., & Krajbich, I. (2019b). Gaze-informed modeling of preference learning and prediction. *Journal of Neuroscience, Psychology, and Economics*, *12*(3–4), 143–158. https://doi.org/10.1037/npe0000107

Smith, S., & Vela, E. (1992). Environmental context-dependent eyewitness recognition. *Applied Cognitive Psychology*, *6*(2), 125–139. https://doi.org/10.1002/(ISSN)1099-0720

Smith, S., & Vela, E. (2001). Environmental and context-dependent memory: A review and meta-analysis. *Psychonomic Bulletin & Review*, *8*(2), 203–220. https://doi.org/10.3758/BF03196157

Sutherland, N., & Mackintosh, N. (1971). *Mechanisms of animal discrimination learning*. Academic Press.

Talluri, B., Urai, A., Tsetsos, K., Usher, M., & Donner, T. (2018). Confirmation bias through selective overweighting of choice-consistent evidence. *Current Biology*, *28*(19), 3128–3135. https://doi.org/10.1016/j.cub.2018.07.052

Tenenbaum, J., & Griffiths, T. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, *24*(4), 629–640. https://doi.org/10.1017/S0140525X01000061

Theeuwes, J. (1992). Perceptual selectivity for color and form. *Perception & Psychophysics*, *51*(6), 599–606. https://doi.org/10.3758/BF03211656

Theeuwes, J. (2010). Top-down and bottom-up control of visual selection. *Acta Psychologica*, *135*(2), 77–99. https://doi.org/10.1016/j.actpsy.2010.02.006

Thomas, A., Molter, F., Krajbich, I., Heekeren, H., & Mohr, P. (2019). Gaze bias differences capture individual choice behavior. *Nature Human Behavior*, *3*(6), 625–635. https://doi.org/10.1038/s41562-019-0584-8

Thura, D., Beauregard-Racine, J., Fradet, C.-W., & Cisek, P. (2012). Decision making by urgency gating: Theory and experimental support. *Journal of Neurophysiology*, *108*(11), 2912–2930. https://doi.org/10.1152/jn.01071.2011

Trueblood, J., Brown, S., & Heathcote, A. (2014). Multiattribute linear ballistic accumulator model of context effects in multialternative choice. *Psychological Review*, *121*(2), 179–205. https://doi.org/10.1037/a0036137

Turner, B. (2019). Toward a common representational framework for adaptation. *Psychological Review*, *126*(5), Article 660. https://doi.org/10.1037/rev0000148

Turner, B., Kvam, P., Unger, L., Sloutsky, V., Ralston, R., & Blanco, N. (2021). *Cognitive inertia: How loops among attention, representation,*

and decision making distort reality. https://doi.org/10.31234/osf.io/8zvey

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*(4), 327–352. https://doi.org/10.1037/0033-295X.84.4.327

Ullman, S., Vidal-Naquet, M., & Sali, E. (2002). Visual features of intermediate complexity and their use in classification. *Nature Neuroscience*, *5*(7), 682–687. https://doi.org/10.1038/nn870

Usher, M., & McClelland, J. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review*, *108*(3), Article 550. https://doi.org/10.1037/0033-295X.108.3.550

Van Hoeck, N., Ma, N., Ampe, L., Baetens, K., Vandekerckhove, M., & Van Overwalle, F. (2013). Counterfactual thinking: An fMRI study on changing the past for a better future. *Social Cognitive and Affective Neuroscience*, *8*(5), 556–564. https://doi.org/10.1093/scan/nss031

van Moorselaar, D., & Slagter, H. (2020). Inhibition in selective attention. *Annals of the New York Academy of Sciences*, *1464*(1), 204–221. https://doi.org/10.1111/nyas.14304

van Moorselaar, D., Theeuwes, J., & Olivers, C. (2014). In competition for the attentional template: Can multiple items within visual working memory guide attention? *Journal of Experimental Psychology: Human Perception and Performance*, *40*(4), 1450–1464. https://doi.org/10.1037/a0036229

Vanunu, Y., Hotaling, J., Le Pelley, M., & Newell, B. (2021). How top-down and bottom-up attention modulate risky choice. *Proceedings of the National Academy of Sciences*, *118*(39), Article 118. https://doi.org/10.1073/pnas.2025646118

Vapnik, V. (1998). *Statistical learning theory*. Wiley.

Vecera, S., Cosman, J., Vatterott, D., & Roper, Z. (2014). The control of visual attention: Toward a unified account. *Psychology of learning and motivation*, *60*(5), 303–347. https://doi.org/10.1016/B978-0-12-800090-8.00008-1

Wald, A., & Wolfowitz, J. (1948). Optimal character of the sequential probability ratio test. *Annals of Mathematical Statistics*, *19*(3), 326–339. https://doi.org/10.1214/aoms/1177730197

Wason, P., & Johnson-Laird, P. (1972). *Psychology of reasoning: Structure and content* (Vol. 86). Harvard University Press.

White, C., Ratcliff, R., & Starns, J. (2011). Diffusion models of the flanker task: Discrete versus gradual attention selection. *Cognitive Psychology*, *63*(4), 210–238. https://doi.org/10.1016/j.cogpsych.2011.08.001

Yang, S., & Lengyel, M. (2016). Active sensing in the categorization of visual patterns. *eLife*, *5*, Article 66. https://doi.org/10.7554/eLife.12215

Yantis, S., & Egeth, H. (1999). On the distinction between visual salience and stimulus-driven attentional capture. *Journal of Experimental Psychology: Human Perception and Performance*, *25*(3), 661. https://doi.org/10.1037/0096-1523.25.3.661

Yau, Y., Hinault, T., Madeline, T., Cisek, P., Fellows, L., & Dagher, A. (2021). Evidence and urgency related EEG signals during dynamic decision-making in humans. *Journal of Neuroscience*, *41*(26), 5711–5722. https://doi.org/10.1523/JNEUROSCI.2551-20.2021

Zeithamova, D., Dominick, A., & Preston, A. (2012). Hippocampal and ventral medial prefrontal activation during retrieval-mediated learning supports novel inference. *Neuron*, *75*(1), 168–179. https://doi.org/10.1016/j.neuron.2012.05.010

Zeithamova, D., Manthuruthil, C., & Preston, A. (2016). Repetition suppression in the medial temporal lobe and midbrain is altered by event overlap. *Hippocampus*, *26*(11), 1464–1477. https://doi.org/10.1002/hipo.22622

Zeithamova, D., & Preston, A. (2017). Temporal proximity promotes integration of overlapping events. *Journal of Cognitive Neuroscience*, *29*(8), 1311–1323. https://doi.org/10.1162/jocn_a_01116

Appendix A

Derivations

Starting with Equation 2 where evidence for Category c on Trial i is a weighted sum of exemplar activations and their associated category labels, we need to compute the partial derivative of this ratio to specify how the attention vector should change so as to minimize the cross-entropy loss. Although we used vector representation in the main text, our derivation here will show the partial derivative on a dimension-wise basis. Letting $f^{(i)}$ again denote the feedback provided on the ith trial, the partial derivative of the cross-entropy loss is as follows:

$$\frac{\partial}{\partial \alpha_j^{(i)}} \log\left(V_{f^{(i)}}^{(i)}\right) = \frac{\partial}{\partial \alpha_j^{(i)}}\left[\log\left(\sum_n a_n \mathbb{I}\left(f^{(n)} = f^{(i)}\right)\right) - \log\left(\sum_n a_n\right)\right]$$
$$= \frac{1}{\sum_n a_n \mathbb{I}\left(f^{(n)} = f^{(i)}\right)} \frac{\partial}{\partial \alpha_j^{(i)}}\left(\sum_n a_n \mathbb{I}\left(f^{(n)} = f^{(i)}\right)\right) - \frac{1}{\sum_i a_n} \frac{\partial}{\partial \alpha_j^{(i)}}\left(\sum_n a_n\right).$$

Here, the partial derivative operator can be applied linearly to each individual element within the summations, and so we need only compute the derivative of the activation expression in Equation 1 for a single exemplar:

$$\frac{\partial}{\partial \alpha_j^{(i)}} a_n = \frac{\partial}{\partial \alpha_j^{(i)}}\left\{\exp\left(-\delta \sum_{j=1}^{D} \alpha_j^{(i)} |e_j^{(i)} - x_j^{(n)}|\right) m^{(n)}\right\} = \frac{\partial}{\partial \alpha_j^{(i)}}\left\{\prod_{j=1}^{D} \exp\left(-\delta \alpha_j^{(i)} |e_j^{(i)} - x_j^{(n)}|\right) m^{(n)}\right\}$$
$$= m^{(n)} \prod_{k \neq j} \exp\left(-\delta \alpha_k^{(i)} |e_k^{(i)} - x_k^{(n)}|\right) \frac{\partial}{\partial \alpha_j^{(i)}}\left\{\exp\left(-\delta \alpha_j^{(i)} |e_j^{(i)} - x_j^{(n)}|\right)\right\} = m^{(n)} \prod_{j=1}^{D} \exp\left(-\delta \alpha_j^{(i)} |e_j^{(i)} - x_j^{(n)}|\right) \frac{\partial}{\partial \alpha_j^{(i)}}\{-\delta \alpha_j^{(i)} |e_j^{(i)} - x_j^{(n)}|\}$$
$$= -\delta m^{(n)} \prod_{j=1}^{D} \exp\left(-\delta \alpha_j^{(i)} |e_j^{(i)} - x_j^{(n)}|\right) |e_j^{(i)} - x_j^{(n)}|.$$

The partial derivative in Equation 9 can be calculated in a similar manner, where here the feedback associated with Trial i would be replaced with the index corresponding to the leading accumulator at Time t.

(*Appendices continue*)

Appendix B

Notation and Parameters

Table B1
Nomenclature

Symbol	Description
Indices	
i	Trial
j	Dimension
h	Feature
n	Exemplar
t	Within-trial timestep
c	Category label
Task environment	
D	Number of dimensions per stimulus
C	Number of possible category labels
$\mathcal{S}$	Set of dimensions at spatial location
Common elements	
e	True stimulus representation
x	Episodic memory trace
f	Feedback
m	Exemplar memory strength
α	Between-trial attention weight
a	Exemplar activation
V	Category evidence
N	Number of exemplars stored
H	Number of observed features
Additional within-trial elements	
e^*	Working stimulus representation
α^*	Within-trial attention weight
m^*	Exemplar dimension encoding
r	Imputed feature value
z	Imputed feature activation
g	Within-trial gradient update value
L	Dimension fixation prediction (true/false)
Q	Feature encoding status (true/false)

Note. Reference table for the notation used in the technical specifications.

Table B2
Parameter Values

		Case study					
Parameter	Description	1A	1B	2	3	4A	4B
δ_B	Kernel specificity (between trial).	0.05	0.02	1.50	0.01	1.50	1.50
γ_B	Learning rate (between trial)	2.50	2.00	1.50	0.001	0.20	0.20
δ_W	Kernel specificity (within trial)	0.20	0.20	0.35	0.25	0.24	0.24
γ_W	Learning rate (within trial)	0.20	0.20	0.08	0.20	0.17	0.17
κ	Encoding threshold	62	62	28	202	74	74
θ	Feature sampling bias	0.70	0.70	0.80	0.95	0.70	0.70
ϕ	Evidence threshold	0.90	0.90	0.99	0.99	0.90	0.90
σ	Perceptual variability	0.10	0.10	0.15	0.10	0.10	0.10

Note. Table of parameter values that were used to simulate behavioral and eye-tracking data in each case study.

Received June 29, 2021
Revision received April 26, 2022
Accepted May 21, 2022 ■