Distributed Neural Systems Support Flexible Attention Updating during Category Learning

Emily R. Weichart[*], Daniel G. Evans[*],

Matthew Galdo, Giwon Bahg, and Brandon M. Turner[†]

The Ohio State University

---

[*] These authors share first authorship

[†] Corresponding author

Abstract

In order to accurately categorize items, humans learn to selectively attend to stimulus dimensions that are most relevant to the task. Models of category learning describe the interconnected cognitive processes that contribute to attentional tuning as labeled stimuli are progressively observed. The Adaptive Attention Representation Model (AARM), for example, provides an account whereby categorization decisions are based on the perceptual similarity of a new stimulus to stored exemplars, and dimension-wise attention is updated on every trial in the direction of a feedback-based error gradient. As such, attention modulation as described by AARM requires interactions among orienting, visual perception, memory retrieval, prediction error, and goal maintenance in order to facilitate learning across trials. The current study explored the neural bases of attention mechanisms using quantitative predictions from AARM to analyze behavioral and fMRI data collected while participants learned novel categories. GLM analyses revealed patterns of BOLD activation in the parietal cortex (orienting), visual cortex (perception), medial temporal lobe (memory retrieval), basal ganglia (prediction error), and prefrontal cortex (goal maintenance) that covaried with the magnitude of model-predicted attentional tuning. Results are consistent with AARM's specification of attention modulation as a dynamic property of distributed cognitive systems.

Keywords: attention; categorization; learning; model-based fMRI

Distributed Neural Systems Support Flexible Attention Updating during Category Learning

Introduction

When grouping items into categories, humans are extraordinarily adept at identifying regularities across dimensions and mapping features to category labels. As we get to know a new person, for example, we may be able to categorize their mood as happy, sad, or angry based on specific elements of their facial expression, tone of voice, or body language. In an effort to explain how humans can learn new categories quickly even when they are multivariate, probabilistic, or non-linearly separable, computational models of categorization aim to formalize the processing stream that links memories of previous experiences to representations of new items (Galdo, Weichart, Sloutsky, & Turner, 2021; Kruschke, 1992; Love, Medin, & Gureckis, 2004; Nosofsky, 1986). Across contemporary models, dynamic allocation of selective attention to goal-relevant dimensions is often implicated as the critical mechanism through which categorization accuracy improves across trials.

Models differ considerably, however, in their descriptions of how attention is distributed to facilitate categorization accuracy. The influential Generalized Context Model (GCM; Nosofsky, 1986), for example, describes a static distribution of attention based on overall dimension diagnosticity across the items represented in memory. Adaptive attention models, by contrast, suggest that attention is updated on every trial according to a feedback-based error gradient, requiring dynamic monitoring of attention-outcome contingencies (Kruschke, 1992; Love, Medin, & Gureckis, 2004). Although previous fMRI work has provided evidence of representational reorganization in the hippocampus that is consistent with an adaptive attention

account (specifically, SUSTAIN; Mack, Love, & Preston, 2016), questions about the nature of attention, its component processes, and the neural systems that are recruited during attention deployment still remain. The aim of our study, therefore, is to discuss the brain functions that contribute to attentional updating in the context of category learning, and to evaluate a theory of dynamic, gradient-based attention through model-based fMRI analyses.

The current study focuses specifically on the Adaptive Attention Representation Model (AARM; Galdo, Weichart, Sloutsky, & Turner, 2021), an example of the class of adaptive attention models described above. The conceptual basis of AARM comes from context theory, which assumes previously-experienced items (i.e. exemplars) are stored in memory as discrete episodic traces along with associated category labels (Medin & Schaffer, 1978). As in GCM, AARM describes how category representations are formed according to the similarity between new stimuli and stored exemplars. An attention vector weights the influence of plausible feature-to-category mappings when the observer makes a choice. AARM additionally includes mechanisms for feedback-based attention updates, which are intended to optimize future responses with respect to the goals of the learner. AARM's attention updating mechanisms therefore incorporates notions of prediction error in a manner that is conceptually related to models of reinforcement learning (RL). Whereas the equation defining the prediction error signal in standard RL models emerges from a gradient of reward as a function of time (Sutton & Barto, 2018), AARM computes the gradient as a function of attention during each individual trial. In summary, the theory put forth by AARM suggests that attention updating requires dynamic interactions among orienting, visual processing, prediction error, memory encoding and retrieval, and goal-directed behavior.

In previous work, support for AARM's mechanisms of attention allocation was provided by fits to simultaneous streams of choice and eye-tracking data that were collected while participants learned novel categories (Galdo, Weichart, Sloutsky, & Turner, 2021). Across paradigms of varying complexity, AARM accurately predicted increases in accuracy that coincided with increased probability of selectively attending to goal-relevant dimensions, as measured by trial-level gaze fixations. Although these results provided support for AARM by way of eye-tracking data as the terminal output of human attention dynamics (Blair, Watson, Walshe, & Maj, 2009; Rehder & Hoffman, 2005a, 2005b), the extent to which AARM's mechanisms reflect expected patterns of neural activity remains to be determined. The current study therefore investigates the neural plausibility of attention updating as described by AARM, given current knowledge about the multifaceted neural loci of its theoretical subprocesses. In particular, we expect the trial-level magnitude of model-predicted attention updates to covary with blood-oxygen-level-dependent (BOLD) activation in five relevant functional clusters (for review, see Seger & Miller, 2010): 1) parietal cortex (orienting); 2) visual cortex (perceptual processing); 3) hippocampus and medial temporal lobe (episodic memory and recognition); 4) midbrain dopaminergic systems and basal ganglia (prediction error); and 5) prefrontal cortex (PFC; goal maintenance and representation).

For our purposes, we used behavioral and fMRI data that were collected by Mack, Love, and Preston (2016) and were made freely available via the Open Science Foundation (OSF; https://osf.io/5byhb/). In the task, participants were asked to categorize novel insects into two groups according to the features contained in three dimensions: legs, antennae, and mouth. Corrective feedback was provided on every trial, allowing participants to effectively map features to category labels. Given the layers of complexity provided by the task paradigm in the

form of multidimensional stimuli, trial-and-error learning, unidimensional and exclusive-OR (XOR) categorization rules, and rule-switches, we deemed the dataset to be ideal for the purpose of identifying the functional components of adaptive attention.

The current article is organized as follows. We begin by providing a conceptual overview of AARM, highlighting the brain regions hypothesized to contribute to dynamic attentional tuning. Second, we will summarize the methods related to data collection (as described by Mack, Love, & Preston, 2016), model-fitting, and model-based fMRI analyses. Finally, we relate the attentional tuning mechanism in AARM to BOLD activation in the regions of interest (ROIs) identified in our analysis, and discuss our results in terms of canonical category learning findings.

Adaptive Attention Representation Model

Figure 1 provides a conceptual overview of AARM's component mechanisms. Additional mathematical details will be provided in the *AARM Technical Specifications* section to follow. In general, AARM defines the processes through which new items are represented in psychological space and mapped to category labels. Learning (i.e. increased categorization accuracy across trials) is conceptualized as a natural consequence of storing experiences of stimuli and associated feedback as they occur, and preferentially allocating attention to the most relevant dimensions. Here, we will introduce the framework in terms of three core components: Representation, Decision, and Attention (Turner, 2019; Weichart, Galdo, Sloutsky, & Turner, 2021).

The *Representation* component of AARM specifies how the low-level perceptual qualities of a new stimulus are interpreted and contextualized by the observer's goals and

experiences. At the beginning of a trial, attention orients to spatial locations due to a combination of salience and learned relevance. When a new stimulus is introduced, the observer then samples information from dimensions according to a learned trajectory of dimension prioritization. This sampling process activates memories of similar items with known category labels, which allow the observer to form a representation of the stimulus that is relevant to the task. Similarity is determined from the feature-level comparison of the current stimulus to all stored exemplars, and is modulated by attention (Equation 1). As such, an exemplar will be perceived to be more *similar* to the current stimulus if its features match on highly-attended dimensions, or more *dissimilar* if its features mismatch on highly-attended dimensions.

The *Decision* component describes how the observer maps the representation of the current stimulus to a category response. Because corrective feedback is typically provided during category learning tasks, AARM presumes that each stored exemplar carries an association to a known category label. The observer therefore has access to the necessary information for mapping the similarity-based activation of each exemplar to its respective category. As such, the total activation across exemplars that are associated with a common category label can be interpreted as decision evidence in favor of that particular category. When making a response, the observer is presumed to select a category in proportion to the relative decision evidence among the available options (Equation 3).

After the observer makes a decision and corrective feedback is observed, the stimulus and the category label are stored in memory for future use. Within the *Attention* component, AARM subsequently updates attention in a manner that is intended to optimize for the goals of the observer on future trials (e.g. improve accuracy, reduce sampling; Equation 4), and occurs in consideration of the predicted response probability relative to the observed feedback. If a highly-

attended dimension provides evidence in favor of the incorrect category label, for example, attention to that dimension will be reduced. The newly-updated attention vector is fed back into the *Representation* component in preparation for the next trial.

It is critical to highlight that the specifications of the *Representation* and *Decision* components of AARM were based on GCM, a model of categorization that assumes attention is calculated *retrospectively* after all stimuli have been observed (Nosofsky, 1986; Turner, 2019). GCM can generate accurate categorization predictions, given that it uses a stable attention vector that is specified to preferentially consider task-relevant dimensions when making decisions. The GCM conceptualization of attention, however, does not naturally extend to questions of category learning. When in a novel task environment with novel stimuli, the observer cannot possibly know which dimensions are going to be relevant and which to attend unless explicitly instructed. This insight can only come from experience.

AARM's innovation relative to GCM, therefore, lies in its inclusion of a gradient-based mechanism for updating attention according to feedback. Because attention is redistributed on every trial based only on what the observer has experienced up until that point, AARM can account for the gradual accrual of information that is required for identifying the task-relevant dimensions concurrent with learning (Galdo, Weichart, Sloutsky, & Turner, 2021; Weichart, Galdo, Sloutsky, & Turner, 2021).

Relative to other adaptive attention models like ALCOVE (Kruschke, 1992) and SUSTAIN (Love, Medin, & Gureckis, 2004), AARM's advancement is its specification of gradient-based attention updating mechanisms that optimize for the individual goals of the learner, rather than error minimization alone. While further exposition will be provided in the *AARM Technical Specifications* section, the gradient calculation allows for the possibility that

secondary computational goals bear an impact on the representation of new items, such as an implicit desire to maximize information sampling efficiency. Given that it is often the case that multiple dimensions provide similarly diagnostic information, the learner could conceivably seek to reduce time or effort spent on each individual trial by only attending to a subset of informative dimensions before making a response, with minimal detriment to overall accuracy. This idea has been supported by our previous presentation of AARM. When additional mechanisms were added to the model to optimize for secondary computational goals, the expanded variant outperformed a baseline unconstrained variant when fit to behavioral and eye-tracking data (Galdo, Weichart, Sloutsky, & Turner, 2021). While a strict error-reduction policy for attention updating that is standard among contemporary adaptive attention models was sufficient for predicting accuracy across trials, accounting for individualized computational goals in the gradient specification was necessary for predicting trial-level information sampling behavior via eye-tracking. Related mechanisms for dimension reduction have been implemented in RL models as well, and have proven necessary for predicting human-like attention operations in naturalistic multidimensional environments (Niv et al., 2015; Leong et al., 2017).

Hypothesized Neural Systems

As an extension to our previous results, the current study investigates the neural plausibility of AARM's attention updating mechanism. In order for this mechanism to be considered theoretically viable, it should, at a minimum, covary with neural activation in the distributed systems that are hypothesized to contribute to continuous tuning across trials. The neural systems that we expect to be recruited during attentional tuning come directly from the literature on the neural correlates of category and RL. In particular, we discuss five functional

clusters for category learning that were defined by Seger and Miller (2010) in an independent review.

The *parietal cortex* is involved in orientation of spatial attention (Bisley & Goldberg, 2010; Yin et al., 2012), which is instantiated in AARM via the connection between the attention gradient and feature sampling when new stimuli are presented (Point **1** in Figure 1). The *visual cortex* is known to be involved in the formation of low-level perceptual representations (Folstein & Palmeri, 2013; Point **2** in Figure 1). The *hippocampus and medial temporal lobe* are involved in the maintenance and retrieval of past learning instances (Cutsuridis & Yoshida, 2017; O'Reilly & Munakata, 2000; Seger & Miller, 2010), as well as modulation of object representations during category learning (Mack, Love, & Preston, 2016). We therefore expect these regions to be involved in attention modulation in AARM, given the mechanism's critical reliance on activation of past exemplars (Point **3** in Figure 1). The *midbrain dopaminergic systems and basal ganglia* have been implicated in behaviors related to prediction error in RL (Averbeck & O'Doherty, 2021). Because category predictions and observed feedback are critical inputs to the attention updating mechanisms in AARM, we expect model-predicted attention to require the influence of prediction error action selection functions in these regions (Point **4** in Figure 1). The *prefrontal cortex* is known to be involved in goal-directed behaviors, particularly in higher-level monitoring of rule-based performance (Bogdanov, Timmermann, Glaescher, Hummel, & Schwabe, 2018), as would be expected for an update rule that optimizes for the learner's goals of reducing errors and maintaining computational parsimony (Point **5** in Figure 1).

Although we do not make specific predictions about the computations that are performed in each set of brain regions, our study seeks to establish that attentional tuning recruits the

contributions of distributed systems as described by AARM's dynamic structure. Further review of the candidate brain regions and how they relate to category learning are provided in the *Discussion*.

Experimental Methods

Dataset

The task paradigm from Mack et al. (2016) builds upon the classic experiments of Shepard, Hovland, & Jenkins (1961), which have become a benchmark test for models of human category learning. The benchmark study used stimuli that consisted of three binary dimensions to construct six types of category delineations (referred to as Types I-VI). The results, which have been replicated several times (e.g. Crump, McDonnell, & Gureckis, 2013; Nosofsky, Gluck, Palmeri, McKinley, & Glauthier, 1994), showed a progression of learning difficulty from Type I (one dimension was perfectly diagnostic of category membership) to Type VI (all three dimensions needed to be attended to produce a correct response). The observed relative learning rates across category types provide considerable empirical constraint that contemporary theories of category learning are expected to account for in order to be regarded as viable (e.g. Galdo, Weichart, Sloutsky, & Turner, 2021; Goodman, Tenenbaum, Feldman, & Griffiths, 2008; Kruschke, 1992; Nosofsky, Gluck, Palmeri, McKinley, & Glauthier, 1994).

The paradigm designed by Mack et al., (2016) presented participants with three different categorization types within the same task context, using a common set of stimulus features. The paradigm therefore posed a unique challenge to participants, such that they had to identify and adapt to new categorization rules in order to maintain high accuracy. In the original study, the inclusion of rule-switches allowed the authors to investigate the hypothesis that learning in a

dynamic task environment is made possible by continuous modulation of object representations. Model-based fMRI analyses using SUSTAIN (Mack, Love, & Preston, 2016; Love, Medin, & Gureckis, 2004) supported their hypothesis, and provided evidence that shifting attention to rule-relevant dimensions impacted object representations in the hippocampus.

Our study builds upon these results, taking a more general approach to understanding the functional correlates of attention. In particular, we use a latent input approach to analyze whole brain fMRI data, which was described by Turner and colleagues (2017) to be ideal for exploratory analysis. Given that the adaptive attention mechanism specified by AARM requires dynamic interactions among multiple cognitive systems, is there evidence of distributed system coactivation in the brain during attentional tuning? Relevant details of the stimuli and procedures are provided in the following sections, but the reader is directed to Mack et al. (2016) for more information.

Stimuli

Stimuli were 8 images of insects, each of which was comprised of a body, legs, antennae, and a mouth. While all insects had an identical body shape, each of the other dimensions contained one of two possible features: legs could be thick or thin, antennae could be thick or thin, and mouths could be shovel- or pincer-shaped. Participants were instructed to learn how to classify the insects according to their features, using the corrective feedback that would be provided after every trial as a guide. Examples of stimuli are shown in the top panel of Figure 2.

Task Paradigm

As mentioned previously, participants completed three sub-tasks during the experiment, each with a different type of categorization rule (Types I, II, and VI; Shepard, Hovland, &

Jenkins, 1961). From the participants' perspective, sub-tasks were delineated by a change in the instructions. For example, a participant may have been asked to categorize insects according to their temperature preference (warm or cool) during the first sub-task, and according to the hemisphere in which they are typically found (eastern or western) during the second. Beyond the change in instructions, participants were not informed of any potential change in rule complexity.

In the Type I sub-task, the category label of each stimulus could be determined from the feature value of one dimension. For example, participants could learn to selectively attend to the relevant "legs" dimension upon observing that all insects with thick legs preferred warm temperatures and all insects with thin legs preferred cool temperatures. The Type II sub-task used an exclusive disjunction (i.e. XOR) rule, and required participants to attend to two dimensions in order to categorize the insects correctly. Insects typically found in the eastern hemisphere, for example, might have thick antennae with a pincer-shaped mouth *or* thin antennae with a shovel-shaped mouth, whereas insects found in the western hemisphere might have thick antennae with a shovel-shaped mouth *or* thin antennae with a pincer-shaped mouth. In this case, the antennae and mouth dimensions are relevant and the legs dimension is irrelevant. The Type VI sub-task extended the logic of Type II, and required participants to learn the feature-category mappings and contingencies among all three dimensions. As such, all three dimensions were relevant for identifying category membership. All participants completed the Type VI task first, and the subsequent order of Types I and II were counterbalanced between-subjects.

Participants completed the three sub-tasks in the MRI scanner, and indicated category responses using a button box. A sub-task consisted of 4 functional runs, each with 32 trials.

During a trial, the stimulus was presented for a duration of 3.5s, followed by a 0.5-4.5s jittered fixation. Participants were then presented with a feedback screen containing the stimulus, accuracy information, and the correct category label for 2s, followed by a 4-8s jittered fixation. Each functional run lasted 194s and included 4 repetitions of each unique stimulus.

Data Description

The dataset contains MRI and behavioral data from 23 right-handed participants (12 males, age 18-31 years) with normal or corrected-to-normal vision. One participant's data were corrupted and therefore excluded from all analyses presented here. Participants completed 4 consecutive runs corresponding to each of three categorization rules (Types I, II, and VI, as previously described). Out of all data files that were made available by Mack et al. (2016) via OSF, the following were used in the current study: 1) MPRAGE T1 anatomical images (FOV=256mm, 1mm isotropic voxels); 2) 12 functional timeseries acquired with a T2*-weighted multiband EPI sequence (TR=2s, TE=31ms, flip angle = 73 degrees, FOV=220mm, 72 slices, 1.7mm isotropic voxels); and 3) behavioral data consisting of stimulus and timing information, categorization responses, and correct category feedback.

Modeling Procedures

As a complement to the conceptual overview of AARM that was provided previously, we now provide the mathematical details of the model as it was specifically used in our current model-based fMRI analyses. It is worth noting that AARM was originally presented by Galdo et al. (2021) as a general framework that was designed to account for attention "shortcuts" that humans often take when completing a classification task. For example, if stimuli contain a large number of dimensions, adult participants tend to consider only a small subset of them when

making decisions (Blanco, Turner, & Sloutsky, 2021). One interpretation of this behavior is that in addition to the goal of achieving high accuracy on a task, humans simultaneously pursue secondary computational goals like reducing the amount of time and effort they spend on individual trials. The extent to which these shortcuts impact behavior, however, varies according to the demands of the task.

The full AARM framework contains various mechanisms that instantiate biases for computational simplicity. For our current purposes, we used the variant of AARM that was identified in a switchboard analysis conducted by Galdo et al. (2021) to provide the best fits to data across five experiments, including Mack et al. (2016). The model description provided here therefore includes mechanisms for regularization (tendency toward low-dimensional representations) and competition (increasing attention to one dimension results in a decrease in attention to the others). For more information on AARM's mechanisms for attentional shortcuts, the interested reader is directed to Galdo et al. (2021) for a thorough investigation in various contexts of task complexity with quantified comparisons to traditional attention constraints.

AARM Technical Specifications

When introducing model notation, we will use unbolded symbols to represent scalar values, bold lowercase symbols to represent vectors, and bold uppercase symbols to represent matrices.

AARM describes how humans learn to categorize a sequence of stimuli $\mathbf{E} = [\mathbf{e}_1, \mathbf{e}_2, \dots]$. Each D-dimensional stimulus belongs to one of C categories and is represented as row vector $\mathbf{e}_t$, where t denotes the trial number. The model assumes that learning occurs via interactions between two continuously-updated processes: memory acquisition and attention to task-relevant

dimensions. To acquire new memories, the model assumes that the stimulus presented on Trial t, $\mathbf{e}_t$, is stored as an episodic trace $\mathbf{x}_i = [x_{i,1}\ x_{i,2}\ \ldots\ x_{i,D}]^{\mathrm{T}}$ (i.e. an "exemplar"). Each exemplar is associated with a memory strength $m_{t,i}$ and a category label $f_i \in \{1,2,\ldots,C\}$ acquired by feedback. The feature values, memory weights, and category labels associated with the exemplars can be conceptualized as matrices that are updated after each trial is completed. On Trial t, the full history of exemplar feature values are contained within $\mathbf{X}_t = [\mathbf{x}_1\ \ldots\ \mathbf{x}_N]$, memory strengths are contained within $\mathbf{M}_t = [m_{t,1}\ m_{t,2}\ \ldots\ m_{t,N}]$, and the relevant category labels are contained within $\mathbf{F}_t = [f_1\ \ldots\ f_N]$.

When a new stimulus is presented, it activates memories for stored exemplars on the basis of perceived similarity. Similarity is computed by way of a factorizable exponential similarity kernel (Nosofsky, 1986; Shepard, 1987), such that activation $a_{t,i}$ of the i-th exemplar in response to the stimulus $\mathbf{e}_t$ on Trial t is given by

$$a_{t,i} = \exp\left(-\delta \sum_{j=1}^{D} \alpha_{t,j} \left|e_{t,j} - x_{i,j}\right|\right) m_{t,i} \quad (1)$$

where δ is the specificity of the similarity kernel function, and $\alpha_{t,j}$ is the attention applied to the j-th dimension on Trial t. Attention to each dimension can be represented succinctly as a D-dimensional vector $\boldsymbol{\alpha}_t$. The values of $\boldsymbol{\alpha}_t$ modulate the observer's perception of each exemplar's similarity to the current stimulus. For example, in the extreme case where $\alpha_{t,j}$ is 0, the differences across dimension j has no impact on exemplar activation. By contrast, as $\alpha_{t,j}$ approaches infinity, an exemplar must have identical values to the stimulus $\mathbf{e}_t$ along the j-th dimension in order to maintain activation of the exemplar. We account for lag-based memory strength using a modified temporal decay function that allows for different temporal weighting structures depending on three parameters (Pooley, Lee, & Shankle, 2011):

$$m_{t,i} = [1 - (1 - \epsilon_p^i)(1 - \epsilon_r^{N_t - i + 1})](1 - \eta) + \eta, \quad (2)$$

where ϵ_p and $\epsilon_r \in [0,1]$ are primacy and recency weights, $\eta \in [0,1]$ is a lower bound for memory weights, and N_t is the number of exemplars stored on Trial t. After computing each exemplar's activation, a Luce choice rule is used to compute categorization choice probability. Specifically, the probability of making a Category c response is

$$P(''c''|\boldsymbol{\alpha}_t, \mathbf{e}_t, \mathbf{F}_t, \mathbf{X}_t, \mathbf{M}_t) = \frac{\sum_{i=1}^{N} a_{t,i} \mathbb{I}(f_i = c)}{\sum_{i=1}^{N} a_{t,i}}, \quad (3)$$

where $\mathbb{I}(f_i = c)$ is an indicator function that returns a one if the i-th exemplar $\mathbf{x}_i$ is associated with Category c:

$$\mathbb{I}(f_i = c) = \begin{cases} 1 & f_i = c \\ 0 & \text{otherwise} \end{cases}$$

Therefore, the probability of choosing c is the summed similarity of the exemplars associated with the c-th category, normalized by the total activation of all exemplars.

AARM assumes $\boldsymbol{\alpha}_t$ changes according to a competitive stochastic gradient-based update rule in an effort to minimize error, and is subject to attentional constraints of regularization and competition. Although the AARM framework supports other variations of attention update rules (Galdo, Weichart, Sloutsky, & Turner, 2021), the specification that is relevant to the current article is as follows:

$$\boldsymbol{\alpha}_{t+1} = \boldsymbol{\alpha}_t + \boldsymbol{\Gamma}[\nabla_{\boldsymbol{\alpha}} \log(P(f_t|\boldsymbol{\alpha}_t, \mathbf{e}_t, \mathbf{F}_t, \mathbf{X}_t, \mathbf{M}_t)) - \lambda \mathbf{1}], \quad (4)$$

where $\log(P(f_t|\boldsymbol{\alpha}_t, \mathbf{e}_t, \mathbf{F}_t, \mathbf{X}_t, \mathbf{M}_t))$ is the log likelihood of making a choice that is consistent with Feedback f_t on Trial t, and $\mathbf{1}$ is a D-dimensional column vector whose elements are all one.

Here, $\nabla_{\boldsymbol{\alpha}}$ is a shorthand denoting a “gradient operator" for computing the set of partial derivatives of a function $f(\mathbf{a})$ with respect to each element of the vector $\boldsymbol{\alpha} = [\alpha_1, \cdots, \alpha_d]^T$:

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

The positive parameter λ determines the strength of L1-norm or LASSO regularization, and is related to attentional capacity constraints and bias toward low-dimensional representations. $\boldsymbol{\Gamma}$ is a matrix whose diagonal elements contain the gradient step-size parameter γ_0 and off-diagonal elements are $-\beta$ such that

$$\boldsymbol{\Gamma} = \begin{bmatrix} \gamma_0 & -\beta & -\beta & \ldots. & -\beta \\ -\beta & \gamma_0 & -\beta & \ldots. & -\beta \\ -\beta & -\beta & \gamma_0 & \ddots & -\beta \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ -\beta & -\beta & -\beta & \ldots. & \gamma_0 \end{bmatrix},$$

where $\beta, \gamma_0 \in (0, \infty)$. β determines the strength of competition between dimensions during the attention update. In other words, for objective function $g(\boldsymbol{\alpha}_t)$, β controls the extent to which increasing attention to one dimension results in a reciprocal decrease in attention to the other dimensions.

To avoid negative values of attention, $\boldsymbol{\alpha}_t$ is constrained to be positive. However, the attention update equation may still propose negative values. To facilitate unconstrained optimization, attention is updated on the log scale. Setting $\boldsymbol{\upsilon}_t = \log(\boldsymbol{\alpha}_t)$ and using the change-of-variable technique, we can rewrite the attention update equation $\boldsymbol{\upsilon}_t$ as

$$\boldsymbol{\upsilon}_{t+1} = \boldsymbol{\upsilon}_t + \boldsymbol{\Gamma}\left[\left\{\nabla_{\alpha} \log\left(P(f_t|\boldsymbol{\alpha}_t, \mathbf{e}_t, \mathbf{F}_t, \mathbf{X}_t, \mathbf{M}_t)\right) - \lambda \mathbf{1}\right\} \odot \exp(\boldsymbol{\upsilon}_t)\right], \quad (5)$$

where $\odot$ is the element-wise multiplication or Hadamard product operator. Because the logarithm is a one-to-one monotonic function, finding the optimal $\boldsymbol{\upsilon}_t$ is equivalent to finding the optimal $\boldsymbol{\alpha}_t$. Derivations of the attention gradient and a parameter recovery study are provided in Galdo et al. (2021).

Model Fitting

The fits to behavioral data from Mack et al. (2016) that are used in the current study were originally presented by Galdo et al. (2021). The model was fit to data from each participant independently, with the general goal of identifying the set of parameters that maximized the likelihood function provided in Equation 3. In an effort to ensure robust optimization, a three-step algorithmic approach was used. First, a Differential Evolution procedure using the DEoptimR package was implemented for 100 iterations using 13 particles ($2\kappa + 1$, where κ is the number of free parameters) to effectively sample the parameter space and identify reasonable initial values (Brest, Greiner, Boskovic, Mernik, & Zumer, 2006; Storn & Price, 1997). Second, the initial values were used as input in R's base implementation of the Nelder-Mead optimization algorithm (Nelder & Mead, 1965). Third, in the event of failure to meet the base convergence criterion after 1000 iterations, R's base implementation of simulated annealing was used for 5000 iterations (Van Laarhoven & Aarts, 1987). The result of this procedure was a single set of best-fitting parameters for each participant.

A few constraints were imposed in an effort to maintain parameter identifiability. The similarity kernel specificity parameter was constrained to $\delta = 1$ for all participants. Initialized values for the 3-dimensional attention vector $\alpha_0 = [\alpha_{0,1}, \alpha_{0,2}, \alpha_{0,3}]^T$ were constrained to be equivalent such that $\alpha_{0,1} = \alpha_{0,2} = \alpha_{0,3} = \alpha_0^*$, and a single parameter α_0^* was freely estimated. To

initialize the representation, two “background exemplars" per category were provided with feature values of [0.5,0.5,0.5] (Nosofsky, 1986; Turner, 2019). This setting assumes the observer begins the task with equal evidence for each category response, such that the initial state is uncertain rather than uninformed (Estes, 1994). The model contained a total of six free parameters: learning rate (γ_0), initial attention (α_0^*), competition (β), regularization (λ), primacy (ϵ_p), recency (ϵ_r), and baseline memory strength (η).

To facilitate our model-based fMRI analyses, we input each participant’s best-fitting parameters back into the model, along with the corresponding participant’s unique experience of trial-level stimuli and feedback. We were therefore able to generate participant-level predictions for changes in the attention gradient across trials in the Mack et al. (2016) experiment. Because we were interested in observing which brain areas contribute to dynamic changes in attention during learning, we calculated a single “attention gradient magnitude" value for each trial, which was the Euclidean norm of model-generated attention update values: $|\mathbf{u}| = \sqrt{\sum_{j=1}^{D} u_j^2}$, where

$$\mathrm{u} = \boldsymbol{\Gamma}\left[\left\{\nabla_{\alpha}\log\left(P(f_t|\boldsymbol{\alpha}_t, \mathbf{e}_t, \mathbf{F}_t, \mathbf{X}_t, \mathbf{M}_t)\right) - \lambda\mathbf{1}\right\} \odot \exp(\boldsymbol{\upsilon}_t)\right]$$

is the attention update vector shown in Equation 5. The attention gradient magnitude was subsequently used as a regressor in our fMRI analyses.

MRI Data Preprocessing and Analysis

Preprocessing and analysis of the fMRI data was performed primarily using FEAT (fMRI Expert Analysis Tool; Version 6.0.5), a tool within FSL (FMRIB’s Software Library; https://fsl.fmrib.ox.ac.uk/fsl/). Functional EPI data were corrected for excessive motion using MCFLIRT (Jenkinson, Bannister, Brady, & Smith, 2002), stripped of non-brain structures using

BET (Smith, 2002), spatially smoothed with a 3.4mm full-width-half-maximum Gaussian kernel, and temporally filtered with a high pass filter cutoff of 100s. Anatomical T1 images were registered to standard space using FNIRT, which generated a transformation matrix for each participant. To align a participant's functional and anatomical images, the functional data were first registered to the participant's T1 image using the brain-boundary-based registration method in FLIRT (Greve & Fischl, 2009; Jenkinson, Bannister, Brady, & Smith, 2002), then transformed into a standard space (MNI152 with 1mm resolution) by applying the same transformation matrix generated from T1 registration. Additionally, FAST (Zhang, Brady, & Smith, 2001) was used to segment the T1 image into three tissue types: gray matter, white matter, and cerebrospinal fluid (CSF). The CSF mask from this segmentation was subsequently transformed into the functional space in order to extract the timeseries of mean CSF signal from each run.

After preprocessing, we used FSL's general linear model tool (FILM: Woolrich, Ripley, Brady, & Smith, 2001) to conduct a three-level whole-brain generalized linear model (GLM) analysis. The goal was to identify the brain areas involved in attentional tuning, as predicted by AARM. Trial-wise attention gradient magnitudes were generated by AARM, timelocked to the onset of each trial's feedback period, then concatenated to create the regressor of interest.

At the first level of the analysis, a GLM was fit to the timeseries of attention gradient magnitudes in each individual run. The model included 32 trial-specific regressors, which were timelocked to the onset of each stimulus and lasted the duration of the decision period during each trial. These trial-specific regressors were included to ensure that any signal attributed to the attention gradient magnitude was not confounded by the influence of cognitive processes involved in the decision period. Additionally, to isolate the effects of attentional updating from the effects of error processing, trial-level accuracy was included as a regressor during the

feedback periods. The attention gradient magnitude, accuracy, and trial-specific regressors for each of 32 trials were convolved with a standard double-gamma hemodynamic response function, temporally filtered with a high pass filter cutoff of 100s, and prewhitened. The temporal derivatives of these 34 regressors were also included in the GLM. Finally, nuisance regressors representing the standard six motion parameters (pitch, yaw, roll, and x,y,z shifts) and mean CSF signal were added to the model to control for signal which does not originate from the BOLD response. The effect of attentional tuning on BOLD signal was calculated as a contrast of the gradient magnitude regressor versus no activity (i.e., gradient magnitude signal greater than zero).

At the second level of analysis, a fixed-effects model was used to calculate the effect of attentional tuning across all runs within-subject. Because the attentional tuning mechanism in AARM is a general cognitive mechanism that is not constrained by the changing categorization rules of the task, we collapsed across all runs for each participant.

The third level of analysis considered group level effects of attentional tuning. Group effects were identified through a mixed effects GLM, which was fit by FSL's FLAME 1+2 algorithm (Woolrich et al., 2004). The algorithm combines an approximation of the Bayesian posterior distribution and Markov Chain Monte Carlo (MCMC) methods to estimate coefficients for each voxel, and was identified by Eklund and colleagues (2016) to produce minimal false positives (<5%) across a battery of fMRI analyses.

The sample size of N=22 from Mack et al., (2016) was deemed sufficient for our purposes on the basis of three factors: 1) large-scale sensitivity and reliability examinations of group fMRI studies with GLM analyses have indicated that 20 or more participants should be included to achieve sufficient reliability (Thirion, Pinel, Meriaux, Roche, Dehaene, & Poline,

2007; Zandbelt, Gladwin, & Raemaekers et al., 2008); 2) several previous studies using model-based fMRI approaches have identified significant effects during category learning using similar sample sizes (N=18-22; Davis, Love, & Preston, 2012; Mack, Preston, & Love, 2013; Nosofsky, Little, & James, 2012); and 3) recovery of AARM's parameters for fits to individual participants was verified in previous work (Galdo et al., 2021), providing assurance of regressor stability within our core analysis.

Results

We now present our results in two sections. First, we show the behavioral results from Mack et al. (2016) and the corresponding predictions from AARM, including the trajectory of latent attention across trials and rule-changes. Second, we show the results of a model-based fMRI analysis that was designed to identify the brain regions that contribute to attentional tuning, as specified by AARM. Taken together, our results demonstrate that AARM can accurately predict learning in a complex category learning task via a gradient-based attentional tuning signal, and the same signal fluctuates across trials in a manner that is consistent with BOLD activation in regions with known relevance to category learning.

Fits to Behavioral Data

After fitting AARM to data, best-fitting parameters were used to generate a predicted progression of latent attentional tuning and associated responses across trials for each participant. Model-predicted category responses to the unique set of stimuli experienced by each participant were converted to "correct" or "incorrect" accuracy information via comparison to the true category labels. A qualitative evaluation of model fits is shown in Figure 2C, where model-predicted accuracy was aggregated across participants and displayed as an orange line. Observed

group-level mean accuracy is shown as a black line, with a 95% Bayesian credible interval (CI) shown as a gray shaded region. Model predictions fall well within the 95% CI range, and closely follow the trajectory of the group-level mean across trials in both conditions of task order (Left: Task Order 1, Types VI–I–II; Right: Task Order 2, Types VI–II–I). While only qualitative fits are shown here, quantitative comparisons conducted by Galdo et al. (2021) showed that the current model provided the best fits to behavioral data from a set of five studies (including Mack, Love, & Preston, 2016) compared to all alternative specifications of AARM and a selection of competing models.

Figure 2B provides insight into how AARM was able to predict learning across categorization rule types. By updating dimension-wise attention on every trial in response to feedback, AARM gradually learns to prioritize information from the most relevant dimensions. Figure 2B shows an increase in attention that is allocated to the relevant dimensions, as indicated by the corresponding categorization rule type. For example, one group of participants experienced Task Order 1, where Type VI blocks (all three dimensions were relevant) were followed by Type I blocks (one dimension was relevant, two were irrelevant), which were followed by Type II blocks (two dimensions were relevant, one was irrelevant). This information is indicated by the stimuli pictured above Figure 2A, in which the relevant dimensions for each sub-task are highlighted in red. Mapping the relevant dimensions to model-generated attention shown in Figure 2B, we observe that the progression of attention mirrors the prescribed sub-task order. Purple, green, and yellow lines reflecting attention to the legs, antennae, and mouth dimensions respectively all increase during the first sub-task when all three dimensions were relevant for determining category membership. In the second sub-task where only the legs dimension was relevant, the corresponding purple line quickly increases from the starting point,

whereas the green and yellow lines drop off to indicate reduced attention to the antennae and mouth dimensions. In the third sub-task, the antennae and mouth dimensions become relevant, and the legs dimension becomes irrelevant. The green and yellow lines that correspond to the newly-relevant dimensions show an increase in attention relative to the second sub-task, and the purple line decreases. A conceptually similar pattern of predictions was observed for participants who experienced Task Order 2, where the lines representing dimension-wise attention in Figure 2B follow a trajectory that is consistent with dimension relevance in each sub-task.

Figure 2A shows the progression of latent attention gradient magnitude across trials. We observe that the magnitude of between-trial attentional tuning is maximized when choice accuracy is low. As the diagnosticity of dimensions is learned, attention is optimally distributed towards the relevant dimension(s) and therefore, smaller changes of attention are required. Because there is less tuning needed, the gradient magnitude tends to diminish toward zero, but quickly rises again when the categorization rule changes.

Neural Covariation of the Attention Gradient

Trial-level attention gradient magnitude was used as the regressor of interest in our GLM analysis. Correct or incorrect accuracy information was included as an additional regressor in order to isolate changes related to attention from changes specific to error processing. As shown in Figure 2A, the largest magnitude of attentional change tended to coincide with rule-switches. Because AARM uses a cross-entropy loss function to calculate the attention gradient that is highly sensitive to errors, it is well in line with expectation that moments of uncertainty about which dimensions were relevant (Figure 2B) would result in a high probability of predicted errors (orange line, Figure 2C) and correspondingly large adjustments in attention (Figure 2A).

As such, our fMRI GLM analysis was designed to identify ROIs where BOLD activation reflected changes across trials that were consistent with learning and associated changes in attention.

Maps from the group-level GLM were converted to z-scores and were thresholded at $Z \geq 3.1$ within each voxel. Spatially-contiguous voxel clusters were corrected for family-wise error at $p < .001$ (Woo, Krishnan, & Wager, 2014) using FSL's implementation of Gaussian Random Field Theory. Smoothness was estimated using FSL's 'smoothest' function on group level residuals. This resulted in 14 unique clusters where model-generated attention gradient magnitude accounted for significant variability in BOLD signal across trials. Figure 3 shows the spatial location of each ROI in MNI152 standard space. Because some ROIs appear to be non-contiguous when displayed as 2-dimensional slices, each ROI was randomly assigned a unique color to properly visualize the spatial differentiation. Sagittal and axial slices in Figure 3 were selected in an effort to display all ROIs as parsimoniously as possible. Table 1 shows the corresponding MNI coordinates and peak Z value of each ROI, where ROIs are listed in descending order of cluster size.

We observe a high degree of overlap between the ROIs identified here, and the five functional clusters of interest as defined by Seger and Miller (2010). The largest ROI (ROI 1 in Table 1) is primarily reflective of the *parietal cortex* and *visual cortex* functional clusters, which are thought to be used for spatial orientation and low-level perceptual object representations during category learning, respectively. The *hippocampus and medial temporal lobe* functional cluster consists of five ROIs (ROIs 7, 10, 11, 13 and 14 in Table 1), and is thought to form higher-level object representations in reference to previously-encoded stimuli in an effort to orthogonalize experiences in memory. Two ROIs are consistent with the *midbrain dopaminergic*

systems and basal ganglia functional cluster (ROIs 7 and 9 in Table 1), which is thought to be involved in prediction error and converting information inputs into actions. Seven ROIs overlap with the *prefrontal cortex* functional cluster (ROIs 2, 3, 4, 5, 6, 8, and 12 in Table 1), which is involved in action policy updating in the presence of rule-switches and changing environments. In consideration of previous literature, these ROIs characterize a diverse set of neural systems that reflect dynamic adjusting of attentional weights upon observation of feedback, beyond what is accounted for by error processing alone.

Discussion

In the current study, we investigated the hypothesis that adaptive attention mechanisms require the synchronized involvement of orienting, visual processing, memory retrieval, prediction error, and goal maintenance systems in order to effectively facilitate learning of novel categories. Our analytical approach focused specifically on the theoretical predictions of one category learning model, AARM. As illustrated in Figure 1, attention in AARM is influenced by the decision component of the observer's experience on each trial, and is then fed back into the representation component to modulate category activations on subsequent trials. As such, attention is conceptualized as the critical mechanism for learning, while also being an emergent property of the learning process itself. It therefore follows that attentional tuning should engage a diverse distribution of neural systems during category learning that are involved in components of *Representation*, *Decision*, and *Attention* (Figure 1).

In previous work, we demonstrated that AARM can predict human-like learning across several complex category learning paradigms using simultaneous streams of behavioral and eye-tracking data (Galdo, Weichart, Sloutsky, & Turner, 2021). As originally demonstrated by

Rehder & Hoffman (2005a), humans gradually show a fixation preference for the most relevant dimensions over the course of learning tasks, and this fixation bias co-occurs with increasing accuracy. The authors argued that learning is not simply a process of pure stimulus-category association, but rather involves a gradual acquisition of information about dimension relevance that eventually allows the observer to categorize items as efficiently as a model like GCM (Nosofsky, 1986). By fitting AARM to eye-tracking data in previous work, we were able to show that AARM's mechanisms of attention not only predict learning at the level of response accuracy, but at the level of information sampling behaviors as well with increasing reliance on relevant dimensions as the task proceeds. Additional work showed that AARM extends to within-trial dynamics, such that it can accurately predict the order in which individuals will fixate to dimensions after gaining sufficient experience with the structure of the task (Weichart, Galdo, Sloutsky, & Turner, 2021). Because gaze fixations during goal-directed behaviors are often considered to be a terminal output of latent attention processes (Blair, Watson, Walshe, & Maj, 2009; Itti & Koch, 2000; Kuhn, Tatler, & Cole, 2009), demonstrating accurate fixation predictions provided support for AARM's ability to capture how humans interact with new stimuli during learning. The current study took an alternative approach, investigating the dynamic processes that give rise to adaptive attention rather than the behaviors that result from it.

As shown in Figure 2C, AARM predicts changes in accuracy across task blocks that closely resemble the aggregate behavior of human participants: observed behavior and model predictions show a decrease in accuracy after each rule-switch that soon re-approaches ceiling-level performance. Although the available feature values are consistent throughout the task, AARM is able to predict shifts in accuracy by way of feedback-informed attention weights to each dimension (Figure 2B), which naturally incur large update magnitudes immediately

following a rule-switch (Figure 2A). Using attention gradient magnitude as a regressor in a GLM, model-based fMRI analyses identified statistically significant covariation in 14 ROIs. Consistent with our hypothesis, our results provided evidence that latent attention mechanisms in AARM indeed covary with BOLD activation in neural systems canonically involved in orienting, visual perception, memory retrieval, prediction error, and goal maintenance aspects of category learning (Seger & Miller, 2010). We additionally consider our results to be consistent with findings from RL modeling work, in which attention mechanisms are investigated as a vehicle for post-error changes in behavior and neural activation. Niv and colleagues (2015), for example, provided evidence that attentional tuning during an RL paradigm facilitated interactions between the intraparietal sulcus, precuneus, and dorsolateral PFC (dlPFC) to update the task representation and provoke action selection via the basal ganglia. Follow-up work by Leong and colleagues (2017) showed that attention served dual purposes of biasing value computations during the decision period and value-updating across learning, as reflected by activation in the ventromedial PFC (vmPFC) and basal ganglia. Together with the results of the current work, these findings support the notion that attention and learning bear bidirectional influences on one another, in a manner that recruits operations from widely distributed systems across the brain.

While the results presented here provide preliminary neural support for AARM, our approach has several limitations. AARM comprises a set of dynamic mechanisms that are hypothesized to be involved in category learning, but the analyses presented here were not intended to make any claims about the computations that occur in the regions identified. Instead, the interpretations that we can draw from a GLM are limited to the notion that model-generated attention gradient magnitude accounts for significant variability in BOLD signal change in the

regions specified. We additionally opted not to conduct similar analyses with attention signals generated by any alternative theoretical accounts. We therefore do not claim that our results could only be identified by AARM, as it is likely that other adaptive attention models would also recruit activation of similar brain regions. For our purposes, it was sufficient to demonstrate that adaptive attention in AARM covaried with neural activation in a manner that a model with stable attention across trials would not be equipped to do. Finally, it is important to note that the current dataset and analysis cannot suitably arbitrate between activation related to attention updating and activation related to traditional notions of prediction error as described by RL accounts (Sutton & Barto, 2018). This is because 1) prediction error is implicit to AARM's mechanisms for attention updating; and 2) transitions between sub-tasks of the Mack et al. (2016) design naturally give rise to both a high probability of prediction error and the necessity to redistribute attention to newly-relevant dimensions. While we do not consider this distinction to be antithetical to the conclusions presented here, follow-up will investigate AARM's predictions in the context of task paradigms that were designed to dissociate between the respective roles of attention and error processing (e.g. Calderon et al., 2021).

The relative simplicity of our analytical approach nevertheless provided us with the opportunity to explore the potential reach of adaptive attention, without imposing constraints on the particular nature of the connection between the latent signal of interest and neural activation in each region. Now that we have established a set of ROIs that coactivate with attentional tuning, the findings presented here will serve as an impetus for future joint modeling work using AARM as a tool to understand the dynamic neural computations involved in learning (Turner, Forstmann, & Steyvers, 2019; Turner, Forstmann, Love, Palmeri, & van Maanen, 2017; Turner, Forstmann, Wagenmakers, Brown, Sederberg, & Steyvers, 2013). In the following sections, we

discuss the ROIs shown in Figure 3 in terms of the functional clusters for category learning that were defined by Seger and Miller (2010).

Parietal Cortex

The largest ROI that was identified by our GLM analysis contained the superior parietal lobe (ROI 1 in Table 1), which is known to play a role in attention orienting and prioritization (Bisley & Goldberg, 2010). In the context of category learning, the process of tuning attention weights can be understood as a matter of orienting attention to the appropriate dimensions, similar to how attention must reorient following an invalid cue in an attentional cueing task (e.g. Posner cueing paradigm; Posner, 1980). When a spatial location (or object) is cued with an invalid cue, attention to the cued location must be diminished in order to facilitate detection of the target elsewhere, which leads to slower response times on invalid trials (i.e., the cueing effect). In this context, BOLD activation in the superior parietal lobe have been shown to track processing differences between validly- and invalidly-cued targets (Vossel, Weidner, Thiel, & Fink, 2009), and individuals with parietal lesions demonstrate a disrupted ability to inhibit invalid cues (Sapir, Hayes, Henik, Danziger, & Rafal, 2004). Other work has suggested that the lateral intraparietal area (LIP) is critically involved in integrating bottom-up (salience-based) and top-down (relevance-based) influences on overt attention (for review, see Bisley & Goldberg, 2010). In particular, Bisley and Goldberg (2010) argued that LIP serves as a “priority map," whereby saccades occur in proportion to behavioral relevance with influences from rapid visual response. In connection to AARM’s mechanisms for attention, the parietal cortex serves a function that is conceptually consistent with allocation of attention to spatial locations according to a combination of learned dimension relevance with potential influences from secondary computational goals.

Visual Cortex

Along with superior parietal lobe, the largest ROI that we identified also contained the bilateral visual pathways in the visual cortex (ROI 1 in Table 1), which has been shown to be involved in tasks that require visual processing of spatial locations or visual features (Maunsell & Treue, 2006; for review, see Posner & Gilbert, 1999; Ungerleider & Kastner, 2000). Important insights on the role of visual cortex in attention, for example, came from early single-cell recordings from macaques (Chelazzi, Duncan, Miller, & Desimone, 1998; Chelazzi, Miller, Duncan, & Desimone, 2001; Luck, Chelazzi, Hillyard, & Desimone, 1997; McAdams & Maunsell, 1999, 2000), which broadly demonstrated neuronal firing preferences for search targets that closely matched a cue. Some studies have additionally shown that after sufficient training, neurons in the inferior temporal gyrus can selectively respond to targets that match a cue on the basis of a particular, task-relevant feature despite mismatching on others (Bichot, Rossi, & Desimone, 2005; De Baene, Ons, Wagemans, & Vogels, 2008; Sigala & Logothetis, 2002) and similar correlates of learned discriminability have been observed via human fMRI (Folstein & Palmeri, 2013; Reber, Gitelman, Parrish, & Mesulam, 2003; Saenz, Buracas, & Boynton, 2002).

In general, the visual cortex is thought to represent objects at the basic perceptual level (e.g. contrast sensitivity and spatial resolution) in a manner that connects to orientation and can be modulated by covert attention (Barbot & Carrasco, 2017; for review, see Carrasco, 2011). It is therefore notable that model-generated attention covaries with low-level sensory processing in the visual cortex.

Hippocampus and Medial Temporal Lobe

Five ROIs overlap with the *hippocampus and medial temporal lobe* functional cluster described by Seger and Miller (2010; ROIs 7, 10, 11, 13 and 14 in Table 1). The medial temporal lobe (MTL) is thought to be responsible for functions related to the encoding and maintenance of individual learning instances (Cutsuridis & Yoshida, 2017; O'Reilly & Munakata, 2000). The CA3 field of the hippocampus is thought to be particularly relevant to category learning, given its role in forming autoassociative links between items. This mechanism is characterized by the representational reactivation of previously-observed items during encoding in order to properly orthogonalize cues that overlap on a subset of dimensions (Becker & Wojtowicz, 2007; Gluck, Meeter, & Myers, 2003; O'Reilly & McClelland, 1994; Sutherland & Rudy, 1989). Learners therefore are able to quickly store activation patterns of similar items with minimal interference (for review, see Hunsaker & Kesner, 2013).

As expected, several studies have demonstrated MTL recruitment during category learning tasks, both alongside human fMRI (Poldrack et al., 2001; Poldrack, Prabhakaran, Seger, & Gabrieli, 1999; Seger & Cincotta, 2006) and monkey neurophysiology methods (Hampson, Pons, Stanford, & Deadwyler, 2004). Other work, however, has suggested that the involvement of the MTL is contingent upon the mode of learning that is required for a particular task. While rule-based categorization (i.e. categories are dissociable by a single dimension) tends to result in maximal differential activation in the hippocampus, information integration (i.e. information from multiple dimensions is required to identify the category) and paradigms that contain unannounced rule-switches tend to additionally recruit the basal ganglia (Poldrack, Prabhakaran, Seger, & Gabrieli, 1999; Seger & Cincotta, 2005) and prefrontal cortex (Nomura et al., 2007; Nomura & Reber, 2008).

The MTL is nevertheless consistently recruited during initial training across paradigms (Poldrack et al., 2001; Poldrack, Prabhakaran, Seger, & Gabrieli, 1999). This suggests that the MTL is necessary for learning, but that familiarity-based activation may be insufficient for categorization in more complex tasks. Studies have shown that item representations in the hippocampus are reorganized in accordance with changing rule states when multiple training periods occur within a single experiment (Aly & Turk-Browne, 2016a, 2016b). Importantly, model-based fMRI work using SUSTAIN additionally showed that this reorganization is influenced by selective attention to dimensions with learned relevance to the current task state (Mack, Love, & Preston, 2016). In light of these results as well as the fact that attention updating in AARM critically relies on continuous comparisons of probes to stored exemplars, identifying ROIs in the MTL that covary with model-predicted attention was consistent with expectation.

Midbrain Dopaminergic Systems and the Basal Ganglia

Two ROIs overlap with the *midbrain dopaminergic systems and basal ganglia* functional cluster (ROIs 7 and 9 in Table 1). The basal ganglia are thought to serve as a hub for converting information inputs to actions, in the form of selecting both movements (Humphries, Stewart, & Gurney, 2006) and task strategies (Frank, 2005). As part of the midbrain dopaminergic system (Schultz & Romo, 1992), their role in action selection is critically influenced by reward-related influxes in dopamine (Schultz, Apicella, Ljungberg, Romo, & Scarnati, 1993; Seymour, Daw, Dayan, Singer, & Dolan, 2007). The superior colliculus, for example, has been shown to be involved in RL by way of biasing visual responses in a reward-seeking manner (Shires, Joshi, & Basso, 2010).

Model-based RL accounts explain that this type of learning can arise from the continuous calculation of prediction errors, which are the differences between expected and observed rewards following particular sequences of actions (Nasser, Calu, Schoenbaum, & Sharpe, 2017; Schultz, 2016; Frank & Badre, 2012a/b). More generally, RL comprises an iterative process of prediction, action selection, observation of outcome, and error-based policy (i.e. strategy) updating, such that observers use their experiences to guide future behaviors. While seemingly straightforward, RL implicitly raises the problem of balancing exploration and exploitation: is it better to exploit an action that is already known to produce a reward, or to explore other actions in the hopes of acquiring a larger, less effortful, or more consistent reward? A compelling line of computational and neurophysiology work (Frank, Doll, Oas-Terpstra, & Moreno, 2009; Frank, Moustafa, Haughey, Curran, & Hutchinson, 2007; Frank, Seeberger, & O'Reilly, 2004; Humphries, Khamassi, & Gurney, 2012) has suggested that the explore vs. exploit tradeoff is directly modulated by striatal dopamine, such that increasing tonic striatal dopamine decreases the probability of explorative action selection output from the basal ganglia to the superior colliculus. fMRI studies have additionally shown that exploration tends to engage the frontal pole whereas exploitation engages the vmPFC (Daw, O'Doherty, Dayan, Seymour, & Dolan, 2006), suggesting dissociable downstream executive effects of action selection via the basal ganglia (Averbeck & O'Doherty, 2021).

In the context of category learning, the basal ganglia are involved in tasks that require learning by trial and error (Cincotta & Seger, 2007). Similar to action selection in RL, it has been suggested that the basal ganglia are involved in the selection of category representations and strategies for sampling information from various dimensions (Seger, 2008; Seger & Miller, 2010) with the goal of maximizing accuracy. Turner et al. (2021), for example, provided

evidence that observers may "exploit" dimensions via fixations that are known to carry probabilistic category information, or they may "explore" other dimensions in the hopes of identifying the one that is most reliably diagnostic of category membership. ROI results are consistent with the expectation that model-generated attention covaries with activation related to prediction error and policy updating in these regions.

Prefrontal Cortex

Seven ROIs overlap with the *prefrontal cortex* functional cluster (ROIs 2, 3, 4, 5, 6, 8, and 12 in Table 1). The PFC is broadly thought to be involved in goal-directed behavior (for review, see Bogdanov, Timmermann, Glaescher, Hummel, & Schwabe, 2018). In category learning tasks where the goal is to efficiently discriminate between categories, goal-directed behaviors refer to the rapid identification and exploitation of the categorization rule. Evidence from monkey neurophysiology has shown robust learning-related differences in neuronal firing between categories, even when stimuli contain multiple overlapping irrelevant features (Freedman, Riesenbuber, Poggio, & Miller, 2002, 2003; Freedman, Riesenhuber, Poggio, & Miller, 2001). Similarly, human fMRI work has shown that learned boundaries between categories as well as relevant feature conjunctions in information integration tasks are represented in the PFC (Jiang, Bradley, & Rini, 2007; Li, Mayhew, & Kourtzi, 2009).

The PFC has been shown to engage during category learning (Reber, Stark, & Squire, 1998; Vogels, Sary, Dupont, & Orban, 2002), and PFC activation is the earliest predictor of the choice after category distinctions have been acquired (Antzoulatos & Miller, 2011, 2014; Djurfeldt, Eleberg, & Graybiel, 2001; Pasupathy & Miller, 2005). The PFC has additionally been shown to be involved in error monitoring and corrective behaviors, particularly in the anterior

cingulate cortex (ACC) and dlPFC (Antzoulatos & Miller, 2014; Carter et al., 1998; Hadland, Rushworth, Gaffan, & Passingham, 2003).

While the basal ganglia appear to be involved in tuning the current stimulus-action policy from trial to trial, the PFC is responsible for higher-level monitoring to identify rule-shifts and inhibit the newly-ineffective policy as needed (Bissonette, Powell, & Roesch, 2013). Interactions between the ACC and dlPFC have therefore been frequently identified in tasks that involve set-shifting, like the Wisconsin Card Sorting Task (Monchi, Petrides, Petre, Worsley, & Dagher, 2001). Because AARM predicts attention updates in the direction of an error gradient, it is consistent with expectation that the increased error frequency that accompanied rule-shifts were associated with both substantial changes to the distribution of attention and increased activity in the PFC.

Conclusions

AARM defines a mechanism of attentional tuning that arises as a consequence of the observer's categorization decisions in relation to feedback, and in turn, directly impacts the psychological representations of future stimuli. Therefore, attention is adaptive in that it adjusts to the experiences of the individual, and facilitates learning in a goal-directed manner. The current study demonstrated that with its unique specification of attentional tuning, AARM was able to accurately predict behavior in a complex task paradigm that required continuous monitoring of goals and representations. Importantly, the attentional tuning mechanisms that made it possible for AARM to predict human-like learning behaviors also covaried with activation in distributed neural systems that have been implicated in distinct aspects of category learning. Given that learning is known to require complex interactions among cognitive

functions of orienting, visual perception, memory retrieval, prediction error, and goal maintenance, our results provide preliminary support for AARM as a neurally-plausible theory for how these interactions occur, and are facilitated by continuous updates to attention.

Data Availability Statement

Data were collected by Mack, Love, and Preston (2016) and are freely available via the Open Science Foundation (OSF; https://osf.io/5byhb/). Model code will be available upon publication at https://github.com/MbCN-Lab.

Acknowledgements

This work was supported by a CAREER award from the National Science Foundation (BMT).

Authors' Contributions

ERW: Writing-Original Draft, Writing-Review & Editing; **DGE:** Formal analysis, Visualization, Writing-Original Draft, Writing-Review & Editing; **MG:** Formal analysis; **GB:** Validation, Writing-Review & Editing; **BMT:** Conceptualization, Funding acquisition, Methodology, Project administration, Supervision, Writing-Review & Editing

References

Aly, M., & Turk-Browne, N. (2016a). Attention promotes episodic encoding by stabilizing hippocampal representations. *Proceedings of the National Academy of Sciences*, *113*(4), E420–E429. DOI: https://doi.org/10.1016/j.neubiorev.2007.07.006.

Aly, M., & Turk-Browne, N. (2016b). Attention stabilizes representations in the human hippocampus. *Cerebral Cortex*, *26*(2), 783–796. DOI: https://doi.org/10.1093/cercor/bhv041.

Antzoulatos, E., & Miller, E. (2011). Differences between neural activity in prefrontal cortex and striatum during learning of novel abstract categories. *Neuron*, *71*(2), 243–249. DOI: https://doi.org/10.1016/j.neuron.2011.05.040.

Antzoulatos, E., & Miller, E. (2014). Increases in functional connectivity between prefrontal cortex and striatum during category learning. *Neuron*, *83*(1), 216–225. DOI: https://doi.org/10.1016/j.neuron.2014.05.005.

Averbeck, B., & O'Doherty, J. (2021). Reinforcement-learning in fronto-striatal circuits. *Neuropsychopharmacology*, 1–16. DOI: https://doi.org/10.1038/s41386-021-01108-0.

Barbot, A., & Carrasco, M. (2017). Attention modifies spatial resolution according to task demands. *Psychological Science*, *28*(3), 285–296. DOI: https://doi.org/10.1177/0956797616679634.

Becker, S., & Wojtowicz, M. (2007). A model of hippocampal neurogenesis in memory and mood disorders. *Trends in Cognitive Sciences*, *11*(2), 70–76. DOI: https://doi.org/10.1016/j.tics.2006.10.013.

Bichot, N., Rossi, A., & Desimone, R. (2005). Parallel and serial neural mechanisms for visual search in macaque area V4. *Science*, *308*(5721), 529–534. DOI: https://doi.org/10.1126/science.1109676.

Bisley, J., & Goldberg, M. (2010). Attention, inattention, and priority in the parietal lobe. *Annual Review of Neuroscience*, *33*, 1–21. DOI: https://doi.org/10.1146/annurev-neuro-060909-152823.

Bissonette, G., Powell, E., & Roesch, M. (2013). Neural structures underlying set-shifting: Roles of medial prefrontal cortex and anterior cingulate cortex. *Behavioral Brain Research*, *250*, 91–101. DOI: https://doi.org/10.1016/j.bbr.2013.04.037.

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. DOI: https://doi.org/10.1037/a0016272.

Blanco, N., Turner, B., & Sloutsky, V. (2021). *The benefits of immature cognitive control: How distributed attention guards against learning traps*.

Bogdanov, M., Timmermann, J., Glaescher, J., Hummel, F., & Schwabe, L. (2018). Causal role of the inferolateral prefrontal cortex in balancing goal-directed and habitual control of behavior. *Scientific Reports*, *8*(1), 1–11. DOI: https://doi.org/10.1038/s41598-018-27678-6.

Braunlich, K., & Love, B. (2019). Occipitotemporal representations reflect individual differences in conceptual knowledge. *Journal of Experimental Psychology: General*, *148*(7), 1192. DOI: https://doi.org/10.1037/xge0000501.

Brest, J., Greiner, S., Boskovic, B., Mernik, M., & Zumer, V. (2006). Self-adapting control parameters in differential evolution: A comparative study on numerical benchmark problems. *IEEE Transactions on Evolutionary Computation*, *10*, 646–657. DOI: https://doi.org/10.1109/TEVC.2006.872133.

Calderon, C., De Loof, E., Ergo, E., Snoeck, A., Boehler, C., & Verguts, T. (2021). Signed reward prediction errors in the ventral striatum drive episodic memory. *Journal of Neuroscience, 41(8),* 1716-1726. DOI: https://doi.org/10.1523/JNEUROSCI.1785-20.2020.

Carrasco, M. (2011). Visual attention: The past 25 years. *Vision Research*, *51*(13), 1484–1525. DOI: https://doi.org/10.1016/j.visres.2011.04.012.

Carter, C., Braver, T., Barch, D., Botvinick, M., Noll, D., & Cohen, J. (1998). Anterior cingulate cortex, error detection, and the online monitoring of performance. *Science*, *280*(5364), 747–749. DOI: https://doi.org/10.1126/science.280.5364.747.

Chelazzi, L., Duncan, J., Miller, E., & Desimone, R. (1998). Responses of neurons in inferior temporal cortex during memory-guided visual search. *Journal of Neurophysiology*, *80*(6), 2918–2940. DOI: https://doi.org/10.1152/jn.1998.80.6.2918.

Chelazzi, L., Miller, E., Duncan, J., & Desimone, R. (2001). Responses of neurons in macaque area V4 during memory-guided visual search. *Cerebral Cortex*, *11*(8), 761–772. DOI: https://doi.org/10.1093/cercor/11.8.761.

Cincotta, C., & Seger, C. (2007). Dissociation between striatal regions while learning to categorize via feedback and via observation. *Journal of Cognitive Neuroscience*, *19*(2), 249–265. DOI: https://doi.org/10.1162/jocn.2007.19.2.249.

Crump, M., McDonnell, J., & Gureckis, T. (2013). Evaluating Amazon's mechanical turk as a tool for experimental behavioral research. *PloS One*, *8*(3), e57410. DOI: https://doi.org/10.1371/journal.pone.0057410.

Cutsuridis, V., & Yoshida, M. (2017). Memory processes in medial temporal lobe: Experimental, theoretical and computational approaches. *Frontiers in Systems Neuroscience*, *11*, 19. DOI: https://doi.org/10.3389/fnsys.2017.00019.

Daw, N., O'Doherty, J., Dayan, P., Seymour, B., & Dolan, R. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, *441*(7095), 876–879. DOI: https://doi.org/10.1038/nature04766.

Davis, T., Love, B., & Preston, A. (2012). Learning the exception to the rule: Model-based fMRI reveals specialized representations for surprising category members. *Cerebral Cortex, 22(2),* 260-273. DOI: https://doi.org/10.1093/cercor/bhr036.

De Baene, W., Ons, B., Wagemans, J., & Vogels, R. (2008). Effects of category learning on the stimulus selectivity of macaque inferior temporal neurons. *Learning & Memory*, *15*(9), 717–727. DOI: https://doi.org/10.1101/lm.1040508.

Djurfeldt, M., Eleberg, O., & Graybiel, A. (2001). Cortex-basal ganglia interaction and attractor states. *Neurocomputing*, *38*, 573–579. DOI: https://doi.org/10.1016/S0925-2312(01)00413-1.

Eger, E., Henson, R., Driver, J., & Dolan, R. (2007). Mechanisms of top-down facilitation in perception of visual objects studied by fMRI. *Cerebral Cortex*, *17*(9), 2123–2133. DOI: https://doi.org/10.1093/cercor/bhl119.

Eklund, A., Nichols, T., & Knutsson, H. (2016). Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proceedings of the National Academy of Sciences, 113(28).* 7900-7905. DOI: https://doi.org/10.1073/pnas.1602413113.

Estes, W. (1994). *Classification and cognition*. Oxford University Press. DOI: https://doi.org/10.1093/acprof:oso/9780195073355.001.0001.

Folstein, J., & Palmeri, T. (2013). Category learning increases discriminability of relevant object dimensions in visual cortex. *Cerebral Cortex*, *23*(4), 814–823. DOI: https://doi.org/10.1093/cercor/bhs067.

Frank, M. (2005). Dynamic dopamine modulation in the basal ganglia: A neurocomputational account of cognitive deficits in medicated and nonmedicated parkinsonism. *Journal of Cognitive Neuroscience*, *17*(1), 51–72. DOI: https://doi.org/10.1162/0898929052880093.

Frank, M. & Badre, B. (2012a). Mechanisms of hierarchical reinforcement learning in cortico-striatal circuits 1: Computational analysis. *Cerebral Cortex, 22(3),* 509-526. DOI: https://doi.org/10.1093/cercor/bhr114.

Frank, M. & Badre, B. (2012b). Mechanisms of hierarchical reinforcement learning in cortico-striatal circuits 2: Evidence from fMRI. *Cerebral Cortex, 22(3),* 527-536. DOI: https://doi.org/10.1093/cercor/bhr117.

Frank, M., Doll, B., Oas-Terpstra, J., & Moreno, F. (2009). The neurogenetics of exploration and exploitation: Prefrontal and striatal dopaminergic components. *Nature Neuroscience*, *12*(8), 1062. DOI: https://doi.org/10.1038/nn.2342.

Frank, M., Moustafa, A., Haughey, H., Curran, T., & Hutchinson, K. (2007). Genetic triple dissociation reveals multiple roles for dopamine in reinforcement learning. *Proceedings of the National Academy of Sciences*, *104*, 16311–16316. DOI: https://doi.org/10.1073/pnas.0706111104.

Frank, M., Seeberger, L., & O'Reilly, R. (2004). By carrot or by stick: Cognitive reinforcement learning in parkinsonism. *Science*, *306*, 1940–1943. DOI: https://doi.org/10.1126/science.1102941.

Freedman, D., Riesenbuber, M., Poggio, T., & Miller, E. (2002). Visual categorization and the primate prefrontal cortex: Neurophysiology and behavior. *Journal of Neurophysiology*, *88*(2), 929–941. DOI: https://doi.org/10.1152/jn.2002.88.2.929.

Freedman, D., Riesenbuber, M., Poggio, T., & Miller, E. (2003). A comparison of primate prefrontal and inferior temporal cortices during visual categorization. *Journal of Neuroscience*, *23*(12), 5235–5246. DOI: https://doi.org/10.1523/JNEUROSCI.23-12-05235.2003.

Freedman, D., Riesenhuber, M., Poggio, T., & Miller, E. (2001). Categorical representation of visual stimuli in the primate prefrontal cortex. *Science*, *291*(5502), 312–316. DOI: https://doi.org/10.1126/science.291.5502.312.

Galdo, B., Weichart, E., Sloutsky, V., & Turner, B. (2021). The quest for simplicity in human learning. DOI: https://doi.org/10.31234/osf.io/xgfmb.

Gluck, M., Meeter, M., & Myers, C. (2003). Computational models of the hippocampal region: Linking incremental learning and episodic memory. *Trends in Cognitive Sciences*, *7*(6), 269–276. DOI: https://doi.org/10.1016/S1364-6613(03)00105-0.

Goodman, N., Tenenbaum, J., Feldman, J., & Griffiths, T. (2008). A rational analysis of rule-based concept learning. *Cognitive Science*, 108–154. DOI: https://doi.org/10.1080/03640210701802071.

Greve, D., & Fischl, B. (2009). Accurate and robust brain image alignment using boundary-based registration. *Neuroimage*, *48*(1), 63–72. DOI: https://doi.org/10.1016/j.neuroimage.2009.06.060.

Hadland, K., Rushworth, M., Gaffan, D., & Passingham, R. (2003). The anterior cingulate and reward-guided selection of actions. *Journal of Neurophysiology*, *89*(2), 1161–1164. DOI: 10.1152/jn.00634.2002. DOI: https://doi.org/10.1073/pnas.0400162101.

Hampson, R., Pons, T., Stanford, T., & Deadwyler, S. (2004). Categorization in the monkey hippocampus: A possible mechanism for encoding information into memory. *Proceedings of the National Academy of Sciences*, *101*, 3184–3189. DOI: https://doi.org/10.1073/pnas.0400162101.

Humphries, M., Khamassi, M., & Gurney, K. (2012). Dopaminergic control of the exploration-exploitation trade-off via the basal ganglia. *Frontiers in Neuroscience*, *6*, 9.

Humphries, M., Stewart, R., & Gurney, K. (2006). A physiologically plausible model of action selection and oscillatory activity in the basal ganglia. *Journal of Neuroscience*, *26*(50), 12921–12942. DOI: https://doi.org/10.3389/fnins.2012.00009.

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. DOI: https://doi.org/10.1016/j.neubiorev.2012.09.014.

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, *40*(10-12), 1489–1506. DOI: https://doi.org/10.1016/S0042-6989(99)00163-7.

Jenkinson, M., Bannister, P., Brady, M., & Smith, S. (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage*, *17*(2), 825–841. DOI: https://doi.org/10.1006/nimg.2002.1132.

Jiang, X., Bradley, E., & Rini, R. (2007). Categorization training results in shape- and category-selective human neural plasticity. *Neuron*, *53*, 891–903. DOI: https://doi.org/10.1016/j.neuron.2007.02.015.

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

Kuhn, G., Tatler, B., & Cole, G. (2009). You look where I look! Effect of gaze cues on overt and covert attention in misdirection. *Visual Cognition*, *17*(6-7), 925–944. DOI: https://doi.org/10.1080/13506280902826775.

Leong, Y., Radulescu, A., DeWoskin, V., & Niv, Y. (2017). Dynamic interaction between reinforcement learning and attention in multidimensional environments. *Neuron, 93(2),* 451-463. DOI: https://doi.org/10.1016/j.neuron.2016.12.040.

Li, S., Mayhew, S., & Kourtzi, Z. (2009). Learning shapes the representation of behavioral choice in the human brain. *Neuron*, *62*(3), 441–452. DOI: https://doi.org/10.1016/j.neuron.2009.03.016.

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

Luck, S., Chelazzi, L., Hillyard, S., & Desimone, R. (1997). Neural mechanisms of spatial selective attention in areas V1, V2, and V4 of macaque visual cortex. *Journal of Neurophysiology*, *77*(1), 24–42. DOI: https://doi.org/10.1152/jn.1997.77.1.24.

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

Mack, M., Preston, A., & Love, B. (2013). Decoding the brain's algorithm for categorization from its neural implementation. *Current Biology, 23(30),* 2023-2027. DOI: https://doi.org/10.1016/j.cub.2013.08.035.

Maunsell, J., & Treue, S. (2006). Feature-based attention in visual cortex. *Trends in Neuroscience*, *29*(6), 317–322. DOI: https://doi.org/10.1016/j.tins.2006.04.001.

McAdams, C., & Maunsell, J. (1999). Effects of attention on the reliability of individual neurons in monkey visual cortex. *Neuron*, *23*(4), 765–773. DOI: https://doi.org/10.1016/S0896-6273(01)80034-9.

McAdams, C., & Maunsell, J. (2000). Attention to both space and feature modulates neuronal responses in macaque area V4. *Journal of Neurophysiology*, *83*(3), 1751–1755. DOI: https://doi.org/10.1152/jn.2000.83.3.1751.

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

Monchi, O., Petrides, M., Petre, V., Worsley, K., & Dagher, A. (2001). Wisconsin card sorting revisited: Distinct neural circuits participating in different stages of the task identified by event-related functional magnetic resonance imaging. *Journal of Neuroscience*, *21*(19), 7733–7741. DOI: https://doi.org/10.1523/JNEUROSCI.21-19-07733.2001.

Nasser, H., Calu, D., Schoenbaum, G., & Sharpe, M. (2017). The dopamine prediction error: Contributions to associative models of reward learning. *Frontiers in Psychology*, *8*, 244. DOI: https://doi.org/10.3389/fpsyg.2017.00244.

Nelder, J., & Mead, R. (1965). A simplex method for function minimization. *The Computer Journal*, *7*, 308–313. DOI: https://doi.org/10.1093/comjnl/7.4.308.

Niv, Y., Daniel, R., Geana, A., Gershman, S., Leong, Y., Radulescu, A., & Wilson, R. (2015). Reinforcement learning in multidimensional environments relies on attention

mechanisms. *The Journal of Neuroscience, 35(21),* 8145-8157. DOI: https://doi.org/10.1523/JNEUROSCI.2978-14.2015.

Nomura, E., Maddox, W., Filoteo, J., Ing, A., Gitelman, D., Parrish, T., … Reber, P. (2007). Neural correlates of rule-based and information-integration visual category learning. *Cerebral Cortex*, *17*(1), 37–43. DOI: https://doi.org/10.1093/cercor/bhj122.

Nomura, E., & Reber, P. (2008). A review of medial temporal lobe and caudate contributions to visual category learning. *Neuroscience & Biobehavioral Reviews*, *32*(2), 279–291. DOI: https://doi.org/10.1016/j.neubiorev.2007.07.006.

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

Nosofsky, R., Gluck, M., Palmeri, T., McKinley, S., & Glauthier, P. (1994). Comparing models of rule-based classification learning: A replication and extension of shepard, hovland, and jenkins (1961). *Memory & Cognition*, *22*(3), 352–369. DOI: https://doi.org/10.3758/BF03200862.

Nosofsky, R., 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. DOI: https://doi.org/10.1073/pnas.1111304109.

O'Reilly, R., & McClelland, J. (1994). Hippocampal conjunctive encoding, storage, and recall: Avoiding a trade-off. *Hippocampus*, *4*(6), 661–682. DOI: https://doi.org/10.1002/hipo.450040605.

O'Reilly, R., & Munakata, Y. (2000). *Computational explorations in cognitive neuroscience: Understanding the mind by simulating the brain*. MIT Press.

Pasupathy, A., & Miller, E. (2005). Different time courses of learning-related activity in the prefrontal cortex and striatum. *Nature*, *433*(7028), 873–876. DOI: https://doi.org/10.1038/nature03287.

Poldrack, R., Clark, J., Pare-Blagoev, E., Shohamy, D., Creso, M., Myers, C., & Gluck, M. (2001). Interactive memory systems in the human brain. *Nature*, *414*, 546–550. DOI: https://doi.org/10.1038/35107080.

Poldrack, R., Prabhakaran, V., Seger, C., & Gabrieli, J. (1999). Striatal activation during acquisition of a cognitive skill. *Neuropsychology*, *13*, 564–574. DOI: https://doi.org/10.1037/0894-4105.13.4.564.

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. DOI: https://doi.org/10.1016/j.jmp.2010.08.003.

Posner, M. (1980). Orienting of attention. *Quarterly Journal of Experimental Psychology*, *32*(1), 3–25. DOI: https://doi.org/10.1080/00335558008248231.

Posner, M., & Gilbert, C. (1999). Attention and primary visual cortex. *Proceedings of the National Academy of Sciences*, *96*(6), 2585–2587. DOI: https://doi.org/10.1073/pnas.96.6.2585.

Reber, P., Gitelman, D., Parrish, T., & Mesulam, M. (2003). Dissociating explicit and implicit category knowledge with fMRI. *Journal of Cognitive Neuroscience*, *15*(4), 574–583. DOI: https://doi.org/10.1162/089892903321662958.

Reber, P., Stark, C., & Squire, L. (1998). Cortical areas supporting category learning identified using functional MRI. *Proceedings of the National Academy of Sciences*, *95*(2), 747–750. DOI: https://doi.org/10.1073/pnas.95.2.747.

Rehder, B., & Hoffman, A. (2005a). Eyetracking and selective attention in category learning. *Cognitive Psychology*, *51*(1), 1–41. DOI: 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*, 811–829. DOI: https://doi.org/10.1037/0278-7393.31.5.811.

Saenz, M., Buracas, G., & Boynton, G. (2002). Global effects of feature-based attention in human visual cortex. *Nature Neuroscience*, *5*(7), 631–632. DOI: https://doi.org/10.1038/nn876.

Sapir, A., Hayes, A., Henik, A., Danziger, S., & Rafal, R. (2004). Parietal lobe lesions disrupt saccadic remapping of inhibitory location tagging. *Journal of Cognitive Neuroscience*, *16*(4), 503–509. DOI: https://doi.org/https://doi.org/10.1162/089892904323057245.

Schultz, W. (2016). Dopamine reward prediction-error signalling: A two-component response. *Nature Reviews Neuroscience*, *17*(3), 183–1995. DOI: https://doi.org/10.1038/nrn.2015.26.

Schultz, W., Apicella, P., Ljungberg, T., Romo, R., & Scarnati, E. (1993). Reward-related activity in the monkey striatum and substantia nigra. In A. Arbuthnott & P. Emson (Eds.), *Chemical signalling in the basal ganglia* (Vol. 99, pp. 227–235). Elsevier. DOI: https://doi.org/10.1016/S0079-6123(08)61349-7.

Schultz, W., & Romo, R. (1992). Role of primate basal ganglia and frontal cortex in the internal generation of movements. *Experimental Brain Research*, *91*(3), 363–384. DOI: https://doi.org/10.1007/BF00227834.

Seger, C. (2008). How do the basal ganglia contribute to categorization? Their roles in generalization, response selection, and learning via feedback. *Neuroscience & Biobehavioral Reviews*, *32*(2), 265–278. DOI: https://doi.org/10.1016/j.neubiorev.2007.07.010.

Seger, C., & Cincotta, C. (2005). The roles of the caudate nucleus in human classification learning. *Journal of Neuroscience*, *25*, 2941–2951. DOI: https://doi.org/10.1523/JNEUROSCI.3401-04.2005.

Seger, C., & Cincotta, C. (2006). Dynamics of frontal, striatal, and hippocampal systems during rule learning. *Cerebral Cortex*, *16*(11), 1546–1555. DOI: https://doi.org/10.1093/cercor/bhj092.

Seger, C., & Miller, E. (2010). Category learning in the brain. *Annual Review of Neuroscience*, *33*, 203–219. DOI: https://doi.org/10.1146/annurev.neuro.051508.135546.

Seymour, B., Daw, N., Dayan, P., Singer, T., & Dolan, R. (2007). Differential encoding of losses and gains in the human striatum. *Journal of Neuroscience*, *27*(18), 4826–4831. DOI: https://doi.org/10.1523/JNEUROSCI.0400-07.2007.

Shepard, R. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317–1323. DOI: https://doi.org/10.1126/science.3629243.

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

Shires, J., Joshi, S., & Basso, M. (2010). Shedding new light on the role of the basal ganglia-superior colliculus pathway in eye movements. *Current Opinion in Neurobiology, 20(6).* 717-725. https://doi.org/10.1016/j.conb.2010.08.008.

Sigala, N., & Logothetis, N. (2002). Visual categorization shapes feature selectivity in the primate temporal cortex. *Nature*, *415*(6869), 318–320. DOI: https://doi.org/10.1038/415318a.

Smith, S. (2002). Fast robust automated brain extraction. *Human Brain Mapping*, *17*(3), 143–155. DOI: https://doi.org/10.1002/hbm.10062.

Storn, R., & Price, K. (1997). Differential evolution: A simple and efficient hueristic for global optimization over continuous spaces. *Journal of Global Optimization*, *11*, 341–359. DOI: https://doi.org/10.1023/A:1008202821328.

Sutherland, R., & Rudy, J. (1989). Configural association theory: The role of the hippocampal formation in learning, memory, and amnesia. *Psychobiology*, *17*(2), 129–144. DOI: https://doi.org/10.3758/BF03337828.

Sutton, R., & Barto, A. (2018). Reinforcement Learning: An Introduction. MIT Press.

Thirion, B., Pinel, P., Meriaux, S., Roche, A., Dehaene, S., & Poline, J. (2007). Analysis of a large fMRI cohort: Statistical and methodological issues for group analyses. *NeuroImage, 35(1),* 105-120. DOI: https://doi.org/10,1016/j.neuroimage.2006.11.054.

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

Turner, B., Forstmann, B., Love, B., Palmeri, T., & van Maanen, L. (2017). Approaches to analysis in model-based cognitive neuroscience. *Journal of Mathematical Psychology, 76,* 65-79. DOI: https://doi.org/10.1016/j.jmp.2016.01.001.

Turner, B., Forstmann, B., Wagenmakers, E., Brown, S., Sederberg, P., & Steyvers, M. (2013). A Bayesian framework for simultaneously modeling neural and behavioral data. *NeuroImage, 72*. 193-206.

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. DOI: https://doi.org/10.31234/osf.io/8zvey.

Ungerleider, S., & Kastner, S. (2000). Mechanisms of visual attention in the human cortex. *Annual Review of Neuroscience*, *23*(1), 315–341. DOI: https://doi.org/10.1146/annurev.neuro.23.1.315.

Van Laarhoven, E., & Aarts, E. (1987). Simulated annealing. In *Simulated annealing: Theory and applications* (pp. 7–15). Springer. DOI: https://doi.org/10.1007/978-94-015-7744-1_2.

Vogels, R., Sary, G., Dupont, P., & Orban, G. (2002). Human brain regions involved in visual categorization. *Neuroimage*, *16*(2), 401–414. DOI: https://doi.org/10.1006/nimg.2002.1109.

Vossel, S., Weidner, R., Thiel, C., & Fink, G. (2009). What is “odd" in Posner’s location-cueing paradigm? Neural responses to unexpected location and feature changes compared. *Journal of Cognitive Neuroscience*, *21*(1), 30–41. DOI: https://doi.org/10.1162/jocn.2009.21003.

Weichart, E., Galdo, M., Sloutsky, V., & Turner, B. (2021). As within, so without; as above, so below: Common mechanisms can support between- and within-trial category learning dynamics. DOI: https://doi.org/10.31234/osf.io/94csh.

Woo, C., Krishnan, A., & Wager, T. (2014). Cluster-extent based thresholding in fMRI analyses: Pitfalls and recommendations. *Neuroimage, 91I.* 412-419. DOI: https://doi.org/10.1016/j.neuroimage.2013.12.058.

Woolrich, M., Behrens, T., Beckmann, C., Jenkinson, M., & Smith, S. (2004). Multilevel linear modelling for fMRI group analysis using Bayesian inference. *NeuroImage, 21(4).* 1732-1747. DOI: https://doi.org/10.1016/j.neuroimage.200.3.12.023.

Woolrich, M., Ripley, B., Brady, M., & Smith, S. (2001). Temporal autocorrelation in univariate linear modeling of fMRI data. *Neuroimage*, *14*(6), 1370–1386. DOI: https://doi.org/10.1006/nimg.2001.0931.

Yin, X., Zhao, L., Xu, J., Evans, A., Fan, L., Ge, H., … Liu, S. (2012). Anatomical substrates of the alerting, orienting and executive control components of attention: Focus on the posterior parietal lobe. *PLoS One*, *7*(11), e50590. DOI: https://doi.org/10.1371/journal.pone.0050590.

Zandbelt, B., Gladwin, T., Raemaekers, M., van Buuren, M., Neggers, S., Kahn, R., Ramsey, N., & Vink, M. (2008). Within-subject variation in BOLD-fMRI signal changes across repeated measurements: Quantification and implications for sample size. *Neuroimage, 42(1),* 196-206. DOI: https://doi.org/10.1016/j.neuroimage.2008.04.183.

Zhang, Y., Brady, M., & Smith, S. (2001). Segmentation of brain MR images through a hidden markov random field model and the expectation-maximization algorithm. *IEEE Transactions on Medical Imaging*, *20*(1), 45–57. DOI: https://doi.org/10.1016/j.neuroimage.2009.06.060.

Tables

Region(s)	X	Y	Z	Cluster size	Max Z score
1. Bilateral visual pathways, superior parietal	12	-101	1	117004	11.00
2. Bilateral dorsal ACC, superior frontal gyrus	0	25	45	6283	6.43
3. L middle frontal and precentral gyrus	-48	8	53	4142	6.79
4. R frontal pole	38	53	-5	3985	6.88
5. R superior middle frontal gyrus, premotor cortex	28	-4	49	2556	6.11
6. L superior middle frontal gyrus, premotor cortex	-42	3	62	1432	6.78
7. Thalamus, hippocampus, superior colliculus	-7	-33	-3	1249	5.10
8. R dorsolateral PFC	43	32	34	1212	6.67
9. R insular cortex, putamen, caudate	21	15	0	1124	5.19
10. L posterior middle temporal gyrus	-58	-40	4	1120	6.05
11. R thalamus, parahippocampal gyrus	11	-44	-2	898	5.94
12. L frontal pole	-28	54	9	812	5.70
13. hippocampus	21	-25	-8	747	5.00
14. R posterior middle temporal gyrus	47	-28	-1	392	5.55

Table 1: Regions of interest resulting from fMRI generalized linear model analysis.

Coordinates and clusters are in 1mm MNI152 space. Spatially-contiguous voxel clusters corrected for family-wise error at $p<0.001$. ROIs are listed in descending order of cluster size. L: left; R: right; ACC: anterior cingulate cortex; PFC: prefrontal cortex

Figure Captions

Figure 1: Conceptual overview of the Adaptive Attention Representation Model. *Basic mechanisms that occur within each component during a single trial are shown as a flowchart. Green text indicates information provided to the observer during the trial, and all other processes are considered latent. Red arrows indicate the direct role of the attention gradient. Yellow markers indicate conceptually-associated neural functions. The dotted line indicates that attention modulates the representation of stored exemplars despite not being physically present at the time of stimulus processing. MTL: medial temporal lobe; BG: basal ganglia*

Figure 2: Attention to dimensions affects accuracy. *Circles overlaying the insect stimuli indicate which dimensions were relevant in each sub-task. In all panels, vertical black lines indicate transitions between sub-tasks. (A) Orange lines show mean model-generated gradient magnitude values across participant-level simulations. (B) Purple, green, and yellow lines correspond to mean model-generated attention (α) quantities allocated to leg, antennae, and mouth dimensions, respectively. (C) Lines show means of observed (black) and model-generated (orange) accuracy across participants. Shaded gray regions show the 95% Bayesian posterior credible intervals assuming a Beta(1,1) prior on the probability of responding correctly.*

Figure 3: Regions of interest resulting from fMRI generalized linear model analysis. *Each cluster is presented as a unique color rendered in MNI152 1mm standard space. Arrows in the sagittal slices indicate the position of corresponding axial slices.*