Learning How to Learn: The Interaction Between Attention and Learning as a Mechanism for Dimensionality Reduction in the Brain

Alana Jaskir, Class of 2017
Department of Computer Science
Certificate in Cognitive Science

Advisor: Yael Niv

Abstract

Evidence suggests that attention and learning interact to help extract task-relevant dimensions in a complex environment. How exactly these two mechanisms interplay and craft an internal representation of our environment is still unclear. We tested human participants on a multidimensional task, the Dimensions Task, in which they had to learn through trial and error to maximize reward. We found behavioral evidence suggesting that the learned reward value of a feature influences how that feature is encoded in the brain. Analyses also suggest that participants attend to the whole dimension of higher rewarding features. By drawing on theoretical and other behavioral findings of attention, we present an improved model of human decision-making in the Dimensions Task, which mimics learning in naturalistic task settings. Model comparison provides empirical evidence of informational benefits of selective attention. We hypothesize these benefits, implemented through competitive statistics and gated learning according to the confidence in state representation, aid the brain in simplifying and learning in a high-dimensional world.

This work was funded in part by the Horton Elmer Fund and the ARO grant W911NF-14-1-0101.

Research conducted under IRB Protocol #5291

Acknowledgments

To Yael Niv, for giving me an encouraging, supportive space since freshman year to explore and develop my passion for computational neuroscience. For your guidance and mentorship in this project and beyond. For helping me mature as a scientist and become confident in my abilities as a researcher.

To Nicole Drummond, for our engaging conversations about the Dimensions Task, science, and life at large. For helping me work through challenging concepts and being a sounding board for my inchoate ideas. For your mentorship and friendship.

To Angela Radulescu, for your generous feedback and guidance for this paper. For your constant encouragement and inspiring focus. To the other members of the Dimensions Task crew, Jennifer Bu and Mingbo Cai, for offering their insight.

To Yotam Sagiv, for enabling my obsession with computation neuroscience with constant, nerdy, and impassioned conversation. For graciously rereading this thesis, my presentation slides, and helping me remain calm and driven.

To Hana Lethen and Kristin Gjika, for lovingly correcting my grammar for four years and extending you services to this paper. For your irreplaceable companionship.

To Minseung Choi and Dan Yang, for your unyielding moral support and uplifting attitudes. For pushing me to be a better version of myself. To all those who have cheered me on over the last four years.

And lastly, to Vira Jaskir, Marc Jaskir, and Christa Jaskir, for being my biggest fans and providing me with a support network of love and laughs.

Contents

1 Introduction **5**

1.1 Learning How to Learn: An Interdisciplinary Question 5

1.2 Reward Learning Influences Selective Attention 5

1.3 Selective Attention Aids Reward Learning: A bidirectional relationship 6

1.4 The Dimensions Task 7

1.5 Feature Reinforcement Learning Model with Decay (fRL+Decay) 8

2 First Approach: Preliminary Study and Bottom-up Attentional Effects **10**

2.1 Participants 10

2.2 Task and Procedures 10

2.2.1 Memory as a Proxy for Attention 10

2.3 Behavioral Results 11

2.4 Limitations of Design 12

3 Second Approach: A Redesigned Dimensions Task and Memory Probe **13**

3.1 Participants 13

3.2 Task and Procedures 14

3.3 Behavioral Results 15

3.3.1 RL Model Fitting for Attention Prediction 15

3.3.2 Encoding the Chosen Stimulus 16

3.3.3 Encoding the Unchosen Stimuli 18

3.3.4 Behavioral Conclusions 19

4 Computational Models of Learning and Attention **19**

4.1 Extending the fRL+Decay Model 20

4.1.1 Biasing Choice. 20

4.1.2 Biasing Learning. 23

4.2 The Models 25

4.3 Model Comparison Metric 25

4.4 Model Comparison Results 26

4.5 Predicting Memory Probe 32

4.5.1 Chosen Stimulus 32

4.5.2 Unchosen Stimulus 35

5 Discussion **36**

6 Appendix **40**

A Original Dimensions Task **40**

B Modeling with Bayes' Theorem **40**

C Honor Code **41**

1. Introduction

1.1. Learning How to Learn: An Interdisciplinary Question

We live in a noisy, multidimensional world. Simple daily tasks, such as crossing the street, are accompanied by a bombardment of sensory information that, if unprocessed, makes it difficult to learn and make decisions. While reinforcement learning algorithms have had success in explaining human behavior of simple tasks, these algorithms become inefficient in increasingly complex, naturalistic task settings [Bellman, 1957, Sutton and Barto, 1998]. Yet, our brain appears to handle this 'curse of dimensionality' with ease.

The brain's efficiency in high dimensions is thought to come from its ability to identify structure in its environment. Using this structure, the brain then filters down sensory inputs to internal state representations where relevant information, such as the speed of an oncoming car rather than its color, is emphasized [Gershman and Niv, 2010, Wilson et al., 2014]. This idea parallels those in representation-learning research in machine learning. Representation learning, automating ways to compress big data sets while maintaining the important features, has become a field in itself [Bengio et al., 2013]. How this dynamic feature selection is realized in the brain is still not well known. Understanding how the brain 'learns how to learn' and the computation behind this process can therefore have cross-disciplinary implications.

1.2. Reward Learning Influences Selective Attention

Selective attention, the differential processing of stimuli, allows the brain to give preferential treatment to some stimuli over others. How reward influences selective attention has been thought, until recently, to lie within the classical dichotomy of attention: goal-directed (endogenous) versus stimulus-driven (exogenous) attention. As an example of this dichotomy, imagine driving a car in a friend's neighborhood. The vibrancy of a red stop sign captures your attention due to its salient color (stimulus-driven attention), while street-name signs gain preferential treatment when searching for your friend's street (goal-directed attention). Rewards have been thought to manipulate indirectly selective attention by modulating a subject's motivation to select one stimulus over another in behavioral tasks (that is, to influence endogenous attention) [Anderson, 2015].

However, it has been shown that associative reward directly influences selective attention. Previously rewarded stimuli persistently capture attention even when stimuli are no longer task-relevant or rewarding [Anderson, 2015, Lee and Shomstein, 2014], and [Anderson et al., 2011] demonstrated a distinctly value-driven attentional capture mechanism, disturbing the classical dichotomy. Findings further suggest that it is not solely the associative value of a stimulus that drives attentional capture, but rather the predictability of the stimulus [Lee and Shomstein, 2014, Sali et al., 2014]. Stimuli that are themselves not rewarding but predict information about the rewarding stimulus become prioritized [Le Pelley et al., 2015, Pearson et al., 2015, Bucker et al., 2015, Mine and Saiki, 2015].

1.3. Selective Attention Aids Reward Learning: A bidirectional relationship

The above findings are in line with theoretical work which argues that selective attention has statistical and informational benefits in learning under uncertainty and does not solely arise as a by-product of limited attentional resources [Dayan et al., 2000]. Indeed, as previously discussed, this simplification is crucial for learning in noisy, multidimensional environments. Recent work by Niv et al. has implicated attention in helping subjects learn in a multidimensional task space [Niv et al., 2015], while the orbitofrontal cortex is thought to encode this simplified, abstract representation of the task [Niv et al., 2015, Wilson et al., 2014, Schuck et al., 2016]. This is supported by other behavioral and computational work [Wilson and Niv, 2012, Leong et al., 2017, Jones and Canas, 2010, Marković et al., 2015, Roelfsema and van Ooyen, 2005, Dayan et al., 2000], as well as lesion studies that provide causal evidence for the attention network's function in selecting of stimulus dimensions for successful learning [Vaidya and Fellows, 2016].

All this information combined suggests a two-way interaction between learning and attentional mechanisms. [Leong et al., 2017] offers strong support for a bidirectional relationship between attention and learning, where attention constrains reward associations to relevant dimensions while the brain learns what to attend to through trial and error. Yet, how exactly are dimensions and features of dimensions included and disregarded in this evolving state representation as attention and learning interact? Using the behavior and theoretical literature, can we better model how this representation develops?

To approach this, in this study, participants played a probabilistic multidimensional learning task where, in each game, one feature was associated with a higher probability of reward. Periodically, we explicitly probed participants' working memory of an entire dimension in order to understand how learned feature values affect attention across trials. By first using the fRL+Decay model from [Niv et al., 2015], it was found that if subjects — according to the model — perceived a feature (e.g. pink) to have higher value, they had a higher probability of recalling that feature in the memory probe. That is, the value of a feature modulated the degree to which that feature was encoded in the brain. Furthermore, evidence was found to support the hypothesis that overtly attending to a feature (e.g. pink) affects how the dimension of that feature (e.g. color) is processed. This paper will examine these behavioral results, their implications on information processing in the brain, and then present a series of computational models that offer descriptions of the brain's underlying algorithms.

1.4. The Dimensions Task

The "Dimensions Task," a multi-armed bandit task, offers a way to study the interaction between attention and learning in a probabilistic-reward setting that simulates real-world conditions [Wilson and Niv, 2012, Niv et al., 2015]. Variations of this task were used in this paper to uncover how representation learning is computed in the brain. In the original paradigm, on each trial participants were shown three stimuli that varied along three dimensions (e.g., color/shape/texture, example stimuli in Appendix, Figure 11). Each dimension had three features; for example, the color of a stimulus could be red, yellow, or green. These features were randomly permuted among the three stimuli on each trial. During a "game," a single dimension (e.g., color) was chosen to predict reward. One feature in this dimension (e.g., red) was probabilistically rewarded 75% of the time (regardless of shape and texture) while choosing either of the other two features led to reward 25%

of the time. We define this highly rewarding feature, f^*, as the target feature. The participant's goal was to make choices that maximize reward (i.e. learn which feature is the target feature). They were not informed of the rewarding feature or dimension during the games. They were only notified when a new game began, meaning that the most rewarding feature had changed.

1.5. Feature Reinforcement Learning Model with Decay (fRL+Decay)

The feature reinforcement learning (RL) model with decay proposed in [Niv et al., 2015] best described subject choice behavior in the original Dimensions Task in comparison to a spectrum of models ranging from sub-optimal (a naive RL model) to statistically optimal (a Bayesian model for representation). Despite work suggesting that the human brain is Bayes-optimal [Beierholm et al., 2009, Körding and Wolpert, 2004], Niv et al.'s cross-validation model comparisons, as well as interventional, causal model comparisons, suggest that representational learning is not Bayesian in humans [Niv et al., 2015, Geana and Niv, 2014]. It is plausible that the brain does not employ Bayesian methods for representation learning due to Bayes' intractability when scaling to realistic, high dimensional environments.

The fRL+Decay model, rather than tracking the posterior probability of a feature being the target feature given previous choices and rewards (as in the Bayes model), uses reinforcement learning to calculate a weight for each feature. This weight-value for a feature, $W(f)$, represents a prediction of how rewarding selecting a stimulus with that feature will be; this takes advantage of the fact that features, rather than a whole stimulus, are responsible for reward. The model also incorporates a form of memory loss and choice kernel by decaying the values of the features of unchosen stimuli. The fRL+Decay model has three parameters: η (learning rate), β (softmax inverse temperature), and d (decay rate). At the beginning of each game, weights $W(f)$ for each of the nine features are initialized to zero. The model then updates the weight (value) of each feature as the subject makes decisions and receives rewards. The value of each stimulus S is calculated as the sum of its feature weights. After obtaining reward, the weights of each feature of the chosen stimulus are updated as follows:

$$W^{\text{new}}(f) = W^{\text{old}}(f) + \eta\left[R_t - V(S_{\text{chosen}})\right] \quad \forall f \in S_{\text{chosen}} \tag{1}$$

where R_t is the reward at time t (here, 0 or 1). The weights of unchosen stimuli are decayed to zero with a fitted rate d:

$$W^{\text{new}}(f) = (1-d)W^{\text{old}}(f) \quad \forall f \notin S_{\text{chosen}} \tag{2}$$

The model predicts subject's choices according to a softmax function. The sum of features in a stimulus equals the stimulus's value, V, and the probability of the participant choosing the stimulus S_{chosen} becomes:

$$P(\text{choice}) = \frac{e^{\beta * V(S_{\text{chosen}})}}{\Sigma(e^{\beta * V(S_i)})}, \quad i \in \{1,2,3\} \tag{3}$$

When modeling action selection in reinforcement learning algorithms, the inverse temperature β can be thought of as modeling subjects' trade-off between exploiting information about previously rewarding actions and exploring less known actions. With an inverse temperature of zero, the model exhibits random behavior and picks actions uniformly, regardless of action values. As the inverse temperature grows, the selection policy approaches a greedy algorithm; the action with the highest value has the highest probability of selection and the probability of other options scale according to those actions' values [Sutton and Barto, 1998]. In a similar way, the β in the softmax can also indicate, to some degree, the model's confidence in its predictions, assuming that subjects themselves are not behaving randomly. The higher the β, the less noisy the model's prediction.

As shown in [Leong et al., 2017], learning and attention have a bidirectional relationship that is captured loosely in this model by the separate updating of chosen and unchosen features. Learning is restricted to the chosen (strongly attended) stimuli each trial; the decay of unchosen (unattended) stimuli weakens their representation and contribution to the next trial. In initial behavioral analysis, we used this model's feature weights as predictors for behavior in a memory probe. This model was used as a benchmark for comparison with other computational models developed in this paper.

2. First Approach: Preliminary Study and Bottom-up Attentional Effects

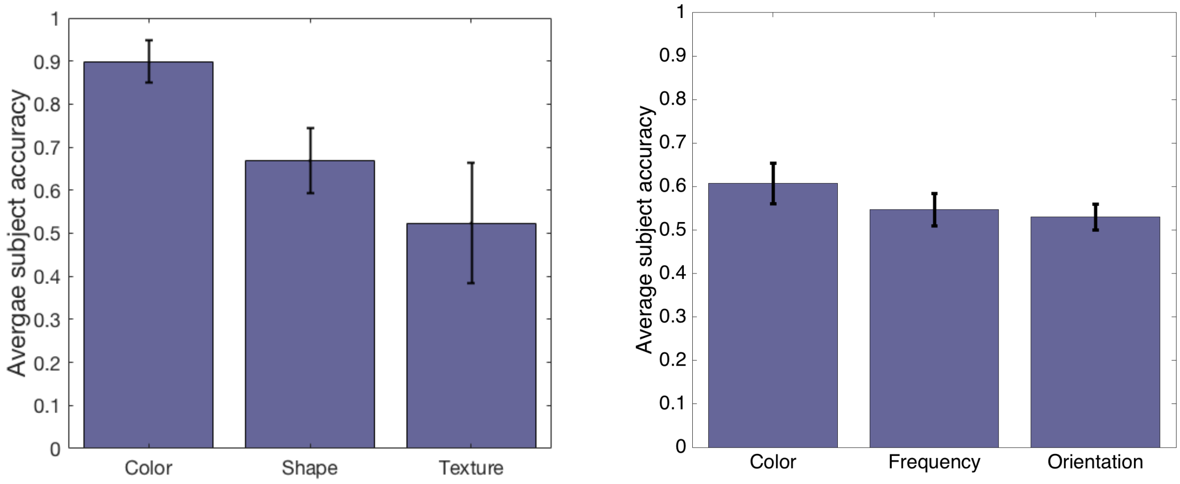
2.1. Participants

Four participants took part in this preliminary pilot. One participant was not included in the analysis due to technical malfunctions during the experiment. Each participant played 60 games. Each game had 20 trials. The experiment lasted approximately an hour.

2.2. Task and Procedures

In order to infer how participants encoded a multidimensional space as a function of learning, participants were to recall the location of a feature from the immediately preceding trial of the Dimensions Task (see Section 1.4 for details). After selecting a stimulus and receiving reward feedback for the choice, three X's appeared on the screen in place of the three stimuli, together with the name of a feature (e.g. red). Participants then selected where they recalled red being in the trial that just ended. This happened at three points during the game: once uniformly between trials 4-7, once uniformly between trials 10-13, and once uniformly between trials 16-19.

2.2.1. Memory as a Proxy for Attention In both approaches presented in this paper, a memory probe is used as a proxy for measuring attention. As previously stated, we are interested in understanding how subjects internally represent a multidimensional space and how reward-learning biases such representation. Specifically, we want to model how attention is distributed to simplify what is learned about in a complex task space and how reward-learning then guides such attention. To measure this, we used the subject's memory of a trial during learning as a snapshot of what the subject gauged to be important for allocating limited resources and for encoding. While the exact relationship between working memory and attention is complex, it is thought that attention acts as a central executive of working memory, controlling what enters working memory and how it is



(a) First Approach, Section 2

(b) Revised Approach, Section 3

Figure 1: Error bars are SEM. (a) Average probe accuracy according to the probed dimension for original Dimensions Task with a single feature probe. (b) Average probe accuracy according to the probed dimension for revised Dimensions Task with a dimensional probe.

maintained [Buschman and Kastner, 2015]. In the case of this experiment, working memory of the stimulus features of a trial gives us insight into how attention during that trial was allocated.

2.3. Behavioral Results

As a reminder, subjects were told to select stimuli which maximize their rewards in the Dimensions Task. When a subject selects the stimulus containing the game's target feature, they are rewarded 75% of the time; else, they are rewarded 25% of the time (see Section 1.4 for full details). A game was considered "learned" if a participant selected the stimulus with the most rewarding feature (the target feature) six times in a row at any point during the game. This happens approximately .1% of the time by chance. Analysis of results showed that there was no significant difference in the proportion of games learned according to dimension (two-tailed t-tests: between color and shape $p > 0.5$, color and texture $p > 0.1$, shape and texture $p > 0.5$). However, despite no statistical difference in learning across dimensions, analysis of probe data revealed a significant difference between dimensions in the average accuracy of probes (Figure 1a). Probe accuracy for color features was significantly higher than average accuracy for shape features and texture features ($ps < 0.05$, one-tailed t-test). There was no significant difference between accuracy to shape and texture. Not only was color accuracy significantly higher than the other dimensions, but its across-subject average was around 90% (95% confidence interval: [0.8495,0.9486]). This suggests that differences in bottom-up salience of the three dimensions - color, shape, and texture - had a large effect on feature memory.

The high recall of color is in line with studies that suggest color may be a more memorable irrelevant dimension [Shin and Ma, 2016]; this may especially be the case when memory of color is in competition with that of shape and texture, more complex dimensions. It could be conjectured that color, particularly the distinct and bold colors used in the stimuli, may need less top-down attention to encode; distinguishing between features is easily resolved with bottom-up attention [McMains and Kastner, 2011]. With participants able to recall color so easily, it is possible that any effects of value learning on encoding may have been greatly mitigated or eliminated.

In order to better measure how learning, rather than bottom-up salience, affects encoding and memory, the Dimensions Task was redesigned to decrease this salience confound as well as other limitations of this paradigm (See Section 2.4). While the revised approach addresses these limitations to better understand nuanced interactions between learning and attention (Section 3), this first approach is important in highlighting how dimensional salience can bias information processing. How this dimensional salience can be manipulated as a function of feature-value learning drives our second approach.

2.4. Limitations of Design

In addition to the significant difference in dimensional recall between color and the other dimensions, other areas of improvement were considered when redesigning the paradigm:

Stimulus Location on Screen. In the original design, the stimuli were presented to the participants in a row across the screen. However, this means that if a subject is selecting and foveating to the stimulus on the left, the visual distance from the participant's focus to the middle and to the right stimuli are not the same. This difference could have some influence on the degree of encoding of the middle versus of the right stimulus. As we want to isolate only the effects of learning on encoding, this feature of the paradigm is undesirable.

Amount of Information from a Probe. In order to get a better understanding of how a participant encodes a multidimensional, multistimulus space, it would be helpful to know how an entire dimension, rather than a single feature, was represented. This would allow us to compare more easily encoding of the stimulus a subject selects (and would presumably contain a valuable feature) to the encoding of features not in that stimulus. It would also allow us to better answer the question of how the degree of attention to a rewarding feature (e.g. red) affects encoding of the entire dimension (i.e. color). While adding complexity to the secondary task may slightly increase distraction from the primary task (i.e. getting reward), the trade-off for better records of working memory would better address the research questions presented.

These limitations were addressed in the revised display and probe of the paradigm presented in Section 3.

3. Second Approach: A Redesigned Dimensions Task and Memory Probe

3.1. Participants

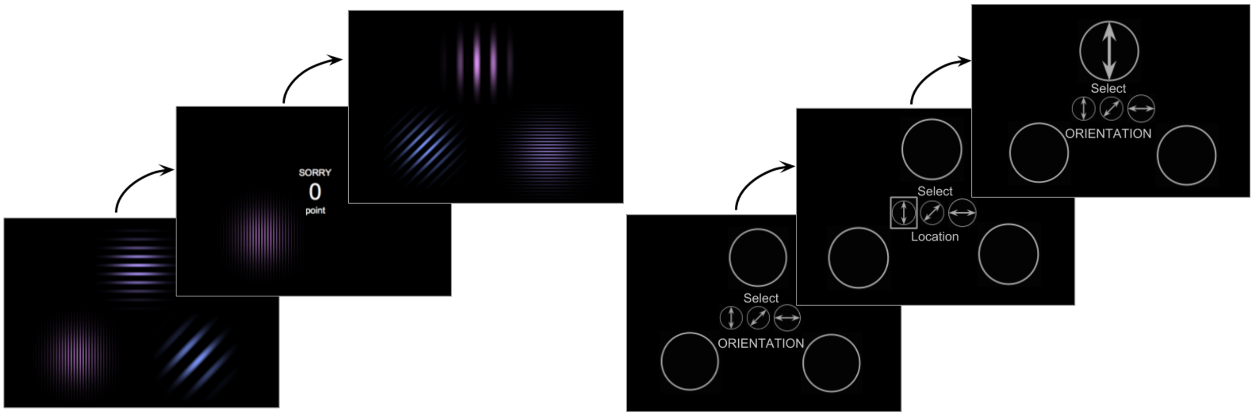
Twenty-four young adults were recruited from the Princeton University community. Three subjects were not used in the analysis due to a preset criterion: missing more than fifty trials and learning fewer than 10% of the games. Learning was defined as correctly selecting the stimulus with the most rewarding feature six times in a row at any point in a game. Each subject played around 54 games. In each game, a subject completed 21 trials. If a trial was missed (a stimulus was not chosen quickly enough), the trial repeated. The experiment lasted approximately one hour.

3.2. Task and Procedures

We used a variant of the Dimensions Task, a multidimensional bandit task, which aimed to equate the salience and low-level visual representability of the dimensions. In this modified task, each stimulus was a Gabor patch that varied in color, orientation, and frequency. These dimensions were chosen so that all dimensions have some mapping to low-level visual processing. This was to counter the effect of dimensional salience on probe memory in approach one (Figure 1a). The three stimuli were presented in a triangle so the visual distance between any two stimuli was the same (Figure 2a).

As in the original paradigm, on each trial participants were shown three stimuli. Each dimension had three features; for example, the color of a stimulus could be pink, purple, or blue. When a subject selects the stimulus containing the game's target feature (selected randomly for each game, e.g. pink) they are rewarded 75% of the time; else, they are rewarded 25% of the time (see Section 1.4 for full details). Participants were told this reward structure in an instructional period. The participant's goal was to select the stimulus on each trial that maximized reward.

Throughout the task, we probed the participant's memory in the beginning, middle, and end of each game (Figure 2b). The probe occurred after the participant selected a stimulus and was given reward feedback for that choice. This probe happened at three points during the game: once uniformly between trials 1-7, once uniformly between trials 8-14, and once uniformly between trials 15-21. The probe consisted of a display of the three features of a dimension. Participants were asked to determine which feature belonged to which stimulus in the immediately preceding trial to the best of their ability. The dimension of the probe was randomly determined. It was emphasized to the participants that they were to play the game as normally as possible, their performance on the probes did not affect their scores, and the occurrence of probes was not related to their performance on the Dimensions Task. This was done to mitigate any influence the probes could have on how participants played the game. The probe allowed us to measure how well a participant had naturally encoded features across a dimension during learning. For a discussion on using a memory probe as a measure of attention, see Section 2.2.1.



(a) The Dimensions Task

(b) Memory Probe

Figure 2: Task Design: (a) Example of a trial: the participant chose a stimulus (pink, vertical orientation, high frequency), was given feedback (0 points, i.e. no reward), followed by three new stimuli for the next trial. (b) Example of a probe: the participant has chosen the "vertical" feature of the orientation dimension and placed it in the location of the top stimulus. The probe continues until the rest of the features are placed.

3.3. Behavioral Results

As a reminder, the redesign of the Dimensions Tasks and memory probe came from the discovery that bottom-up salience may have had some influence on recall of the color dimension, with an average accuracy of a probed color feature around 90% (See Section 2.3). The aim was to reduce the influence of bottom-up salience so as to better measure top-down attentional modulation of learning and state representation. In the new design, the average probe accuracy of the color dimension was still significantly higher than the other two dimensions (frequency $p < 0.005$, orientation $p < 5e-04$) and there was no significance between frequency and orientation ($p > 0.1$). However, the average probe accuracy of color is now around 60% ([0.5642, 0.6598] 95% confidence interval) (Figure 1b); this average memory accuracy was far enough from ceiling performance to allow top-down value learning to influence probe memory, as we hoped to see.

3.3.1. RL Model Fitting for Attention Prediction Participants' behavior was fit to the feature RL model with decay (fRL+Decay) proposed in [Niv et al., 2015] using MATLAB's *fmincon* function (see section 1.5 for details on fRL+Decay). Fitted modeled weights were then used to predict a

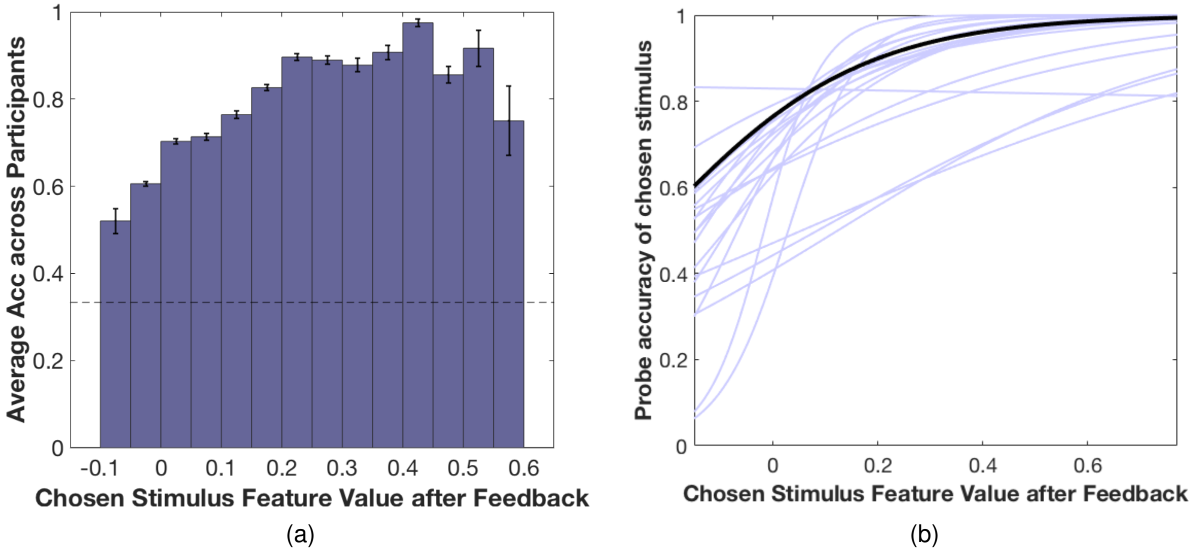


Figure 3: Probe accuracy for the chosen stimulus as a function of its feature weight after feedback, on the probed dimension. The dashed line represents chance. (a). Average accuracies across participants, binned for each participant by the weight of the probed feature of the chosen stimulus. To reduce noise, if a subject had fewer than three data points for a given bin, that bin was not included in the overall average. Error bars are SEM. (b). Logistic regressions of accuracy as a function of modeled feature value. Each lighter line shows the regression of a single subject's data. The dark line is a logistic with average slope and average intercept over all subject regressions.

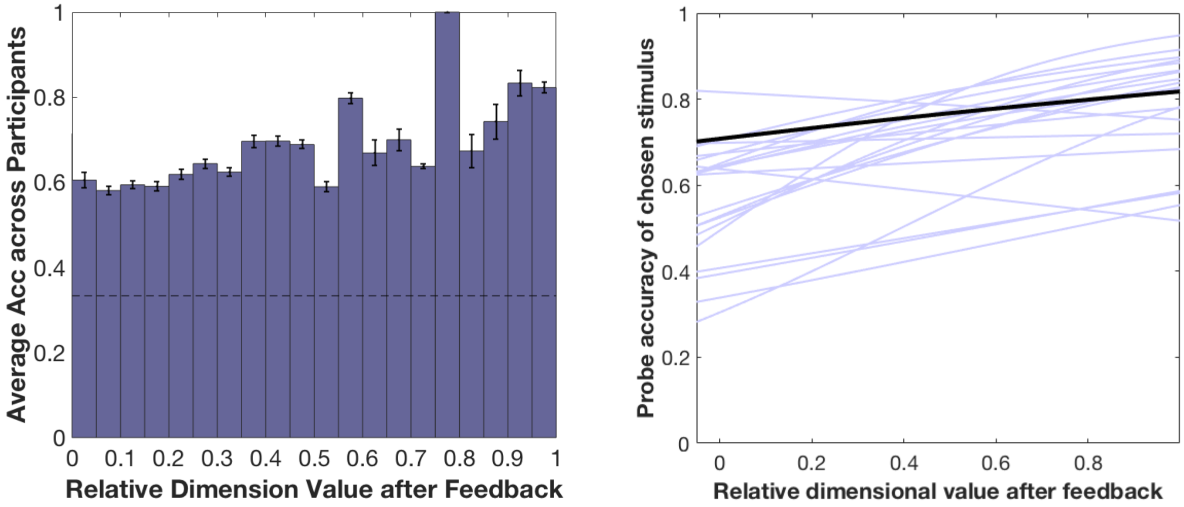
subject's memory of a dimension. We found a difference in the relationship between modeled feature weights and correct performance on the probe test for the chosen stimulus as compared to unchosen stimuli. (As a reminder, the chosen stimulus is the stimulus that the participant selected on the trial just preceding the probe). We therefore report results for chosen and unchosen stimuli separately.

3.3.2. Encoding the Chosen Stimulus When probed, participants in general recalled the feature of the stimulus they chose with above-chance accuracy (Figure 3a), suggesting that they attended to all three features of the stimulus they had selected. Importantly, a random effects analysis showed that the value of the chosen feature (in the probed dimension) significantly affected probe accuracy. For each subject's data, a logistic regression was fit to accuracy on a chosen stimulus, either 1 (correct) or 0 (incorrect), against the modeled value of that feature. A two-tailed t-test showed that the mean of the regression slopes from all subjects was significantly different than zero. The t-statistic specifically indicated a positive relationship between feature value and probe accuracy (t-stat= 6.0027, $p < 1e-05$ when regressing accuracy against modeled values before feedback $W^{\text{old}}(f)$; t-stat= 6.3220, $p < 5e-06$ when regressing accuracy against weights of the features after feedback $W^{\text{new}}(f)$, which depend on $W^{\text{old}}(f)$, Figure 3b). This suggests that although participants attended to all three features of the chosen stimulus, the higher the value of the feature, the more strongly participants encoded that feature in memory. Moreover, the learned weights predicted not only an increase in accuracy for higher positive feature weights, they also predicted a *decrease* in accuracy for features with negative weights, as shown in the histogram in Figure 3a.

A similar trend was found when analyzing a subset of the probe trials in which the feature with the highest modeled weight, f^*, was *not* in the dimension that was probed, D_{probed}. If a dimension did not contain the highest modeled feature weight, $f^* = \max[W(f)]$, its "relative value" was defined as:

$$R_{\text{value}}(D) = \frac{\max[W(f), f \in D]}{f^*} \tag{4}$$

An $R_{\text{value}}(D)$ close to 0 (or less than zero for dimensions with maximum weights that were negative) indicates that the maximum feature weight in the dimension was small in comparison to the maximum feature weight over all dimensions. By contrast, a relative value close to 1 means that the dimension's maximum feature weight is close to f^*. Random effects analysis again showed that the relative value of D_{probed} after feedback showed a positive relationship with the probe accuracy of the chosen stimulus (Figure 4; $p < 1e-05$). That is, as the difference increased between f^*, the maximum weight across dimensions, and the maximum weight of a probed dimension not containing f^*, participants' accuracy on the chosen stimulus decreased. These results suggest that even within the chosen stimulus, not all three dimensions were equally strongly encoded. Rather, features of the chosen stimulus compete for representation (or attention) based on their learned weights (reward-predictive values).



(a)

(b)

Figure 4: Probe accuracy for the chosen stimulus as a function of the probed dimension's relative value. See Equation 4. (a) Average across subject bin averages. To reduce noise, if a subject had fewer than three data points for a given bin, that bin was not included in the overall average. Error bars are SEM (b) Logistic regressions of accuracy against relative dimensional value. Each lighter line shows the regression of a single subject's data. The dark line is a logistic with average slope and average intercept over all subject regressions.

3.3.3. Encoding the Unchosen Stimuli We hypothesized that participants attended to whole dimensions and not just the chosen, highest weighted feature. If this is true, participants should be more accurate in remembering the features of unchosen stimuli when probed on an attended dimension (a dimension with the highest feature weight) than on the unattended ones. To test this, we divided the data into two sets. The first set, P_{max}, was the average subject probe accuracies of unchosen stimuli when f^*, the unique modeled weight maximum before feedback, was in the probed dimension and was in the chosen stimulus. Let P_{other} be the average subject accuracies for when f^* was not in both the probed dimension and chosen stimulus. A one-tailed t-test confirmed that average probe accuracy in P_{max} was higher than in P_{other} $(t_{19} = 2.2331, p < 0.05)$. When separating according to values after feedback, $(t_{19} = 3.1949, p < 0.005)$. This suggests that attention was allocated to the whole dimension and not only to the chosen, highest weighted feature.

3.3.4. Behavioral Conclusions Behavioral analysis suggests that feature values of stimuli learned through a simple reinforcement learning model can be used as predictors of attention and memory in a multidimensional task. We observed a positive relationship between feature weights of stimuli that subjects selected during the task and the recall of those features during an attentional memory probe. This suggests that feature-value learning influences the way in which these features are encoded in an internal representation. Relative value differences between dimensions can also predict encoding. Furthermore, analyses suggest attention is allocated across the whole dimension of the highest feature weight, rather than entirely on a feature-by-feature basis. We now move on to developing computational models that capture this behavior and further explain the interaction between learning and attention in solving multidimensional, probabilistic tasks.

4. Computational Models of Learning and Attention

The above behavioral results suggest that ongoing learning modulates how a multidimensional environment is simplified and stored in the brain. Specifically, value learning biases processing by strengthening the representation of more valuable features and weakening the encoding of less valuable features. This feature-level learning also drives changes in dimensional processing. Participants are more likely to recall unattended stimuli if they belong to the dimension of the highest weighted feature.

The behavioral results of recall accuracy of the chosen stimulus in the task are in line with computational models of learning and attention. The positive relationship between chosen stimulus value and accuracy of recall (Fig 3) supports the model of associative learning which argues that attention to features of a compound stimulus is stronger for features more predictive of reward [Mackintosh, 1975]. An alternative model suggests attention is directed toward features with the most uncertainty, where uncertainty of a feature means there is less evidence or data to accurately predict its value [Pearce and Hall, 1980]. These two views were combined under a model by [Dayan et al., 2000]. A theoretical neural network model, Attention-Gated Reinforcement Learning, shares characteristics with the Dayan et al. (2000) model, specifically competitive dynamics and gated learning [Roelfsema and van Ooyen, 2005]. We will now discuss how these theoretical works can be used to extend the fRL+Decay model by biasing choice and learning. After this, we will combine these extensions to form five different models. We will then compare performance of these models to fRL+Decay.

4.1. Extending the fRL+Decay Model

4.1.1. Biasing Choice. The Dayan et al. (2000) model theorizes that at choice, when estimating a stimulus's prediction of reward, each feature's value should be biased by its relative reliability. That is, the value of a more reliable feature - a feature with a lower variance of its reward history - should contribute more to a combined prediction of reward for a stimulus. The model calculates this relative reliability of a feature, i, compared to all other features, j, at time t in the following way:

$$\pi_i(t) = \frac{\rho_i(t)x_i(t)}{\Sigma_j \rho_j(t)x_j(t)} \tag{5}$$

where $x_j(t) \in \{0,1\}$ (rewarded trial or not) and $\rho_j(t) = 1/\tau_j^2(t)$ is the reliability of stimulus, j. The $\tau_j(t)$ is the standard deviation of each prediction of the value of the feature, $w_j(t)$, around the true value of the feature, j, assuming predictions across trials vary according to a Gaussian distribution around the true feature value. The net prediction based on the features present should then be:

$$\sum_i \pi_i(t)w_i(t) \tag{6}$$

This model incorporates selective attention via statistical competition between features (Equation 5) according to their relative predictability; this relative metric then biases how much a feature's value contributes to a compound stimulus's prediction.

[Leong et al., 2017] found empirically that biasing feature choice with a composite attentional metric (the smoothed product of MVPA and eyetracking) aided fRL performance. (The fRL model calculates stimulus value, selects a stimulus, and updates chosen feature weights the same way as the fRL+Decay model; it does not decay values of features in the unchosen stimulus). In the paper, when calculating the value of a stimulus, rather than directly summing a compound stimulus's features, their model used a sum of the feature weights biased by this composite dimensional-attention measure (Figure 5). This is exactly the structure for the combined net prediction presented in Dayan et al. (2000) in Equation 6.

Feature Attention at Choice. We can extrapolate this competitive dynamics concept and empirical attention findings to bias choice in the fRL+Decay model. Specifically, rather than summing the features of a stimulus to calculate its value, we can bias the feature value by some estimate of relative feature attention, Φ:

$$V(S) = \sum_{f \in S} \Phi(f) W(f) \tag{7}$$

How exactly should Φ be calculated? The Dayan et al. (2000) model uses competition between the reliability of features. However, in the Dimensions Task, subjects know that only one feature, the target feature, is associated with high reward. Therefore, subjects should instead allocate their attention to a feature based on a relative measure of that feature being highly rewarding.

Since only one feature is highly rewarding, this relative metric should exhibit winner-takes-all dynamics. The AGREL model previously mentioned implements this dynamic within a layer of a neural network using the softmax function [Roelfsema and van Ooyen, 2005]. Since a high weight means a feature is more predictive of high reward, we can use our feature weights as inputs to the softmax function. The $\Phi(f)$, our relative feature attention from Equation 7, then becomes:

$$\Phi(f) = \frac{e^{\beta \times W(f)}}{\sum_{i=1}^{9} e^{\beta \times W(f_i)}} \tag{8}$$

If we consider weights to be an evidence accumulator of a feature being the highest rewarding feature, the magnitude of the softmax function's inverse temperature β can then be thought of as how sensitive a subject is to relative differences in evidence amongst the features. (Note that this β is different from the β used to calculate the likelihood of the model predicting a subject's choice). A large β implies only small differences in weights are needed for subjects to exhibit greedy attention

- allocating most of the attention to the highest weighted feature. Conversely, a small β requires more evidence for a subject to disproportionately give one feature attention over others. This logic fits nicely into the explore/exploit dynamic typically used to describe decision-making. In decision-making, the trade-off between exploration and exploitation of different valued choices has been found to have possible neural substrates in the brain [Cohen et al., 2007]. Specifically, explore/exploit behavior is thought to be mediated by the firing mode of norepinephrine neurons in the locus coeruleus (LC). It is plausible that this mechanism could be co-opted in attentional dynamics, where a subject must choose between allocating attention to known rewarding stimuli and allocating attention to less explored stimuli.

Dimension Attention at Choice. Rather than using a relative measure of a feature being the target feature, participants could also bias choice using a relative metric that a dimension contains the target feature. Recall that we used a calculation of relative dimension value to significantly predict memory on the chosen stimulus (Figure 4); the higher the relative dimension value, the more likely subjects were to recall the stimulus. The attentional metric used to improve the fRL model in [Leong et al., 2017] also used dimensional attention (Figure 5). It was found earlier in this paper that feature-level values predict effects of dimensional processing (Section 3.3.3). It is plausible, then, that subjects bias choice using relative dimensional attention. We can again use the softmax function to calculate relative dimensional attention.

First, for each dimension, we took the maximum feature weight in that dimension. This approximates a winner-takes-all dynamic on a dimensional level; these three values were then passed to a softmax to calculate relative dimensional attention. For clarity:

$$V(S) = \sum_{f \in S} \Phi_{D(f)} W(f) \tag{9}$$

where $D(f)$ means the dimension of the desired feature. The dimensional attention, Φ_D, of a dimension, D, is:

$$\Phi_D = \frac{e^{\beta \times \max\{w|w=W(f'),\forall f' \in D\}}}{\sum_{i=1}^{3} e^{\beta \times \max\{w|w=W(f'),\forall f' \in D_i\}}} \tag{10}$$

where D_i is the ith dimension.

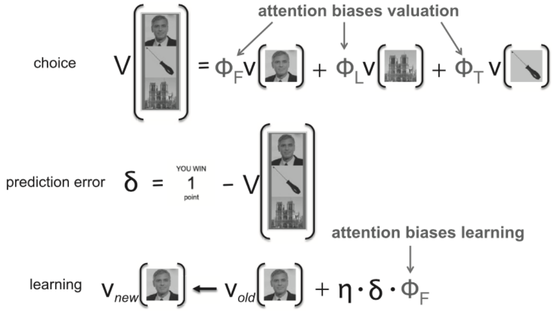

attention biases valuation
choice
V
= Φ_F V + Φ_L V + Φ_T V
prediction error
δ = YOU WIN 1 point − V
attention biases learning
learning
V_new ← V_old + η · δ · Φ_F

Figure 5: Model from [Leong et al., 2017]. The study showed that biasing choice and updating of a fRL model (see 1.5, no decay) with experimental measures of dimensional attention (MVPA and eye-tracking) improved model performance of predicting behavior in a variant of the Dimensions Task.

4.1.2. Biasing Learning. Both the Dayan et al. (2000) model and the Leong et al. model biased learning by some attention metric. The Leong et al. model modulates learning rate using the composite attention metric to the benefit of model performance (Figure 5); the combined learning and choice-biased model outperform models that bias only one or neither parts of the fRL model. The Dayan et al. (2000) model suggests that when associating prediction errors with features, attention and learning should be greater for more uncertain features (i.e. features with low evidence, for example, at the start of a game in the Dimensions Task). As experience is accumulated, uncertainty for some features goes down (e.g. for the feature pink, if the last six stimuli the participant selected were pink) and therefore high learning and attention due to uncertainty should decrease for those features. However, it is important to note that features' uncertainties in the Dimensions Task are dependent. Again, this is due to the structure of the task; since only one feature is highly rewarding, evidence that a feature is the target feature is also evidence that the other features are not highly rewarding. Instead, we again borrow a method from the AGREL model - lowering the learning rate for high contributing features. This makes features with higher certainty and predictability of high reward more resistant to environmental noise.

Let us take the original fRL+Decay update equation, Equation 1:

$$W^{\text{new}}(f) = W^{\text{old}}(f) + \eta\,[R_t - V(S_{\text{chosen}})] \quad \forall f \in S_{\text{chosen}}$$

and modulate a feature's learning rate as a function of its weight:

$$W^{\text{new}}(f) = W^{\text{old}}(f) + \sigma\left(W^{\text{old}}(f)\right)\eta\,[R_t - V(S_{\text{chosen}})] \quad \forall f \in S_{\text{chosen}} \tag{11}$$

where:

$$\sigma\left(W^{\text{old}}(f)\right) = \frac{1}{1+e^{-A(W^{\text{old}}(f))+B}} \tag{12}$$

for which A and B are the slope and intercept of the logistic, respectively. While each feature of the chosen stimulus receives the same global error signal, the added σ allows for a more modular and adaptive learning rate. The σ captures in some sense how [Dayan et al., 2000] theorizes attention at learning should change due to feature uncertainty (low learning for features with high certainty), though feature weights do not perfectly map to the uncertainty of a feature. The sigmoid function has biological feasibility and can be thought of as a gating or threshold neuron that changes activation monotonically according to its input [Bengio et al., 2015]. Model fitting shows that this sigmoid converges to a negative slope, as predicted.

4.2. The Models

By combining the modifications proposed in Section 4.1, we compared five different models:

- *fRL+Decay* As a benchmark of performance, we fit data to the original fRL+Decay model. See Section 1.5.
- *fRL+Decay+Dimensional Attention (DA)* This model implemented a modified calculation of stimulus value by biasing feature weights by a metric of relative dimensional attention. See Equations 9 and 10.
- *fRL+Decay+Feature Attention (FA)* This model implemented the modified calculation of stimulus value by biasing feature weights by a metric of relative feature attention. See Equations 7 and 8.
- *fRL+Decay+DA+Sigmoid Modulated Learning Rate (DAS)* This model implemented fRL+Decay+DA, in addition to fitting a sigmoid which modulated learning rate of a feature according to the weight of that feature. See Equations 11 and 12.
- *fRL+Decay+FA+Sigmoid Modulated Learning Rate (FAS)* This model implemented fRL+Decay+FA, in addition to fitting a sigmoid which modulated learning rate of a feature according to the weight of that feature. See Equations 11 and 12.

All feature weights were initialized to $\frac{1}{9}$. Fitted parameters can be found in Table 1.

4.3. Model Comparison Metric

We compared model probability of the proposed extensions of fRL+Decay by fitting behavioral data from the second approach (Section 3) which used Gabor patches as stimuli. To compare models, we used the Bayesian Information Criterion (BIC) [Schwarz et al., 1978], which is calculated as follows:

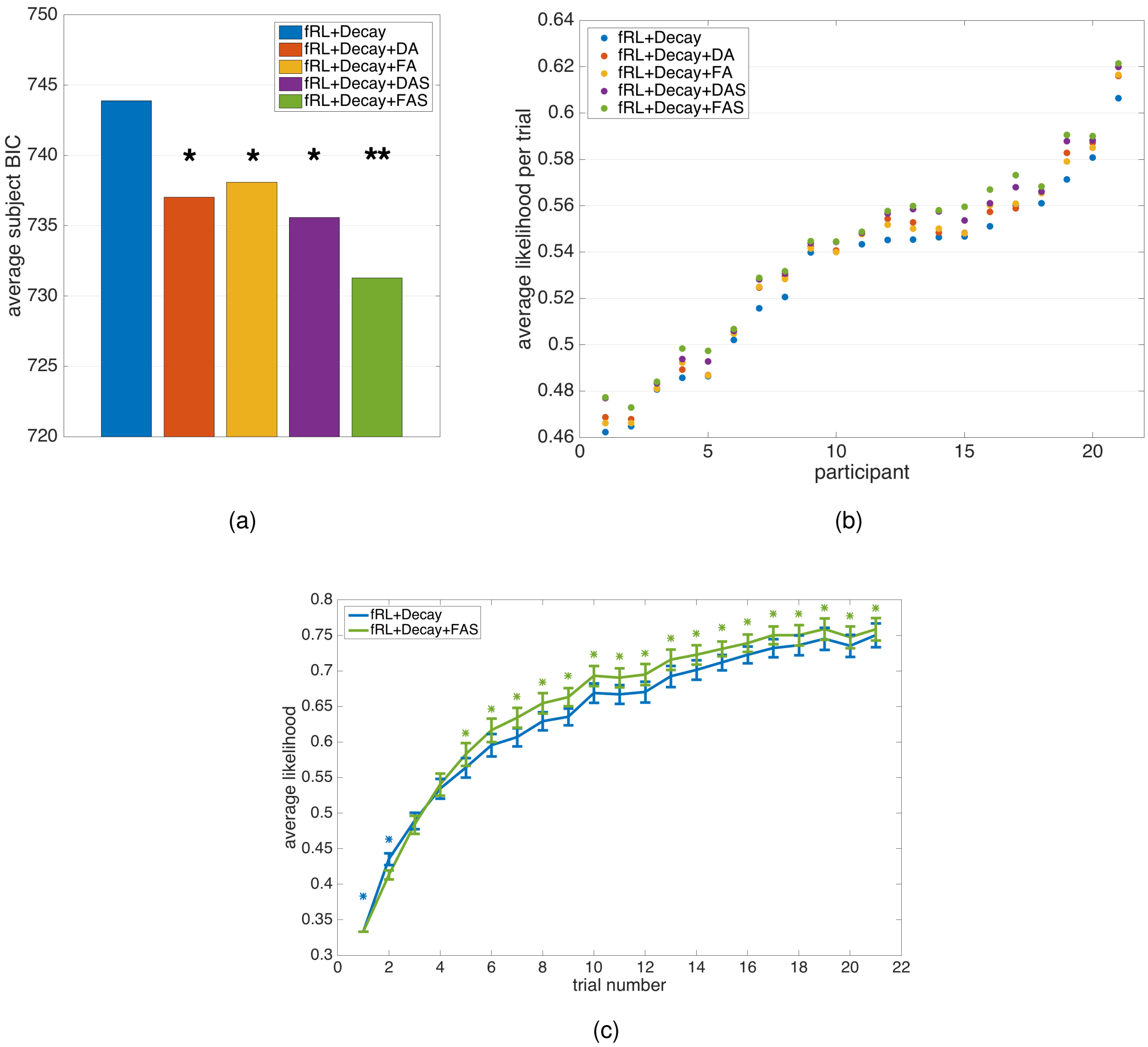
$$BIC_{\text{model}} = -\log\left(P(\text{data}|\text{model})\right) + \frac{\text{M}}{2}\log(\text{N}) \tag{13}$$

where M is the number of parameters in the model and N is the number of data points used to fit the model (product of number of games completed by length of each game). BIC is used to penalize models with high complexity so as to counter high performance with a measure of possible overfitting. Best fitting parameters were found using MATLAB's *fmincon* function. To calculate the likelihood of the data give the model, $\log(P(\text{data}|\text{model}))$, the sum of the log-likelihoods of subject's trial-by-trial choice data was calculated; each trial's log-likelihood was the log of the softmax in Equation 3. The BIC was calculated for each subject individually. The negative of this sum was minimized and used in BIC comparison. The parameters with the lowest negative log-likelihood over ten randomly initialized parameter starting points were used in the comparison. When comparing models, a lower BIC metric signifies a better model.

4.4. Model Comparison Results

Biasing Choice. Biasing choice in the fRL+Decay+DA and fRL+Decay+FA was enough to significantly outperform the reigning fRL+Decay model (BIC subject comparison, two-tailed t-test: $p < 5e-04$, t-stat = 4.15; $p < 1e-03$, tstat $= 3.88$ respectively). The subject BICs between the fRL+Decay+DA and fRL+Decay+FA models were not significantly different (two-tailed t-test, $p > 0.1$). See Figure 6a. These results suggest subjects consider more heavily the values of, and by extension attend more to, features more predictive of high reward or whose dimensions included higher valued features. Furthermore, the competitive dynamic of attention based off relative prediction of reward using a softmax gives empirical evidence for the statistical characteristics of selective attention [Dayan et al., 2000].

Biasing Choice and Learning. The fRL+Decay+FAS outperformed all four other models on the BIC metric. (BIC subject two-tailed t-test results: (fRL+Decay) $p < 5e-06$, t-stat = 6.024; (fRL+Decay+DA) $p < .01$, t-stat $= 3.135$; (fRL+Decay+FA) $p < 1e-03$, t-stat$= 3.944$; (fRL+Decay+DAS) $p < 5e-04$, t-stat = 4.526). While adding a modulated learning rate to the fRL+Decay+DA model maintained its performance over fRL+Decay, its performance was not significantly different than



(a)

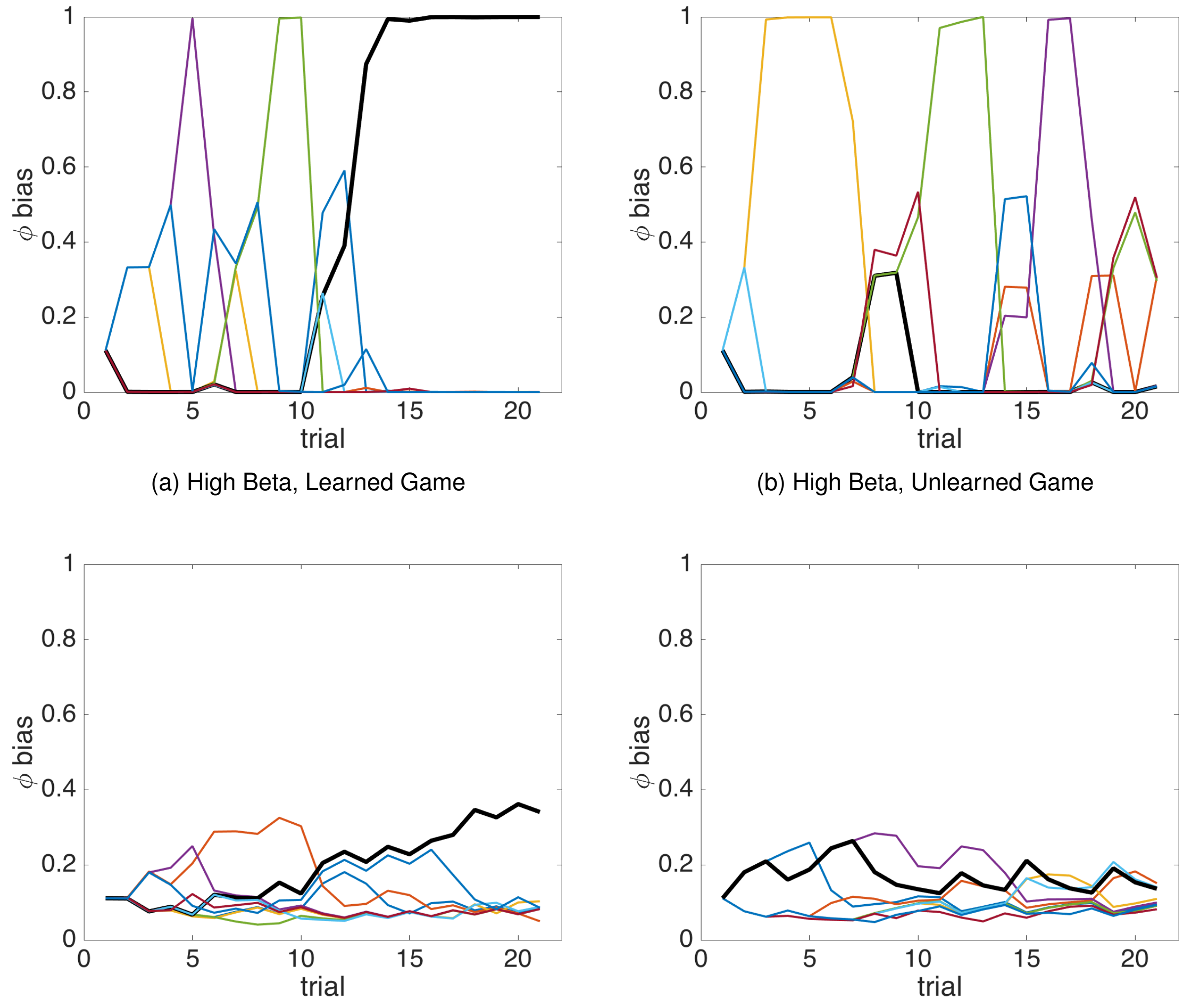
(b)

(c)

Figure 6: Model Comparison. (a) fRL+Decay+FAS outperforms all models on BIC metric, including original fRL+Decay. Other model variants also outperform fRL+Decay but have BICs significantly greater than that of fRL+Decay+FAS. (b) Average likelihood per trial within subjects. (c) Trial average likelihood for fRL+Decay+FAS is significantly above that of fRL+Decay for trials 5-21 (trials 5-18: ps < 5e-05, trials 19-21: ps < .005). fRL+Decay is significantly above fRL+Decay+FAS for trials one and two (p < .005, p < 1e-06). Error bars are SEM.

the model without modulated learning rate or fRL+Decay+FA (Figure 6a). Overall, these results support a predominantly feature-centric rather than dimensional attention during learning.

The high β_2 averages, but large standard deviations, for both fRL+Decay+FAS and fRL+Decay+DAS suggest individual subject preferences on how much to trust small weight differences when biasing attention at choice (Table 1). Despite the large range of β_2 values, the extremes of these β_2 values



(a) High Beta, Learned Game

(b) High Beta, Unlearned Game

(c) Low Beta, Learned Game

(d) Low Beta, Unlearned Game

Figure 7: Example of how relative feature attention Φ (used to bias choice in fRL+Decay+FAS) fluctuations in a game. Each line is the relative feature attention for a specific feature; the black line is the relative feature attention of the game's most rewarding feature. (a-b) Two example games from subject with highest β_2 value, used to calculate the Φ bias. $\beta_2 = 88.0299$ (c-d) Two example games from subject with lowest $\beta_2 = 1.064$ value, used to calculate the Φ bias.

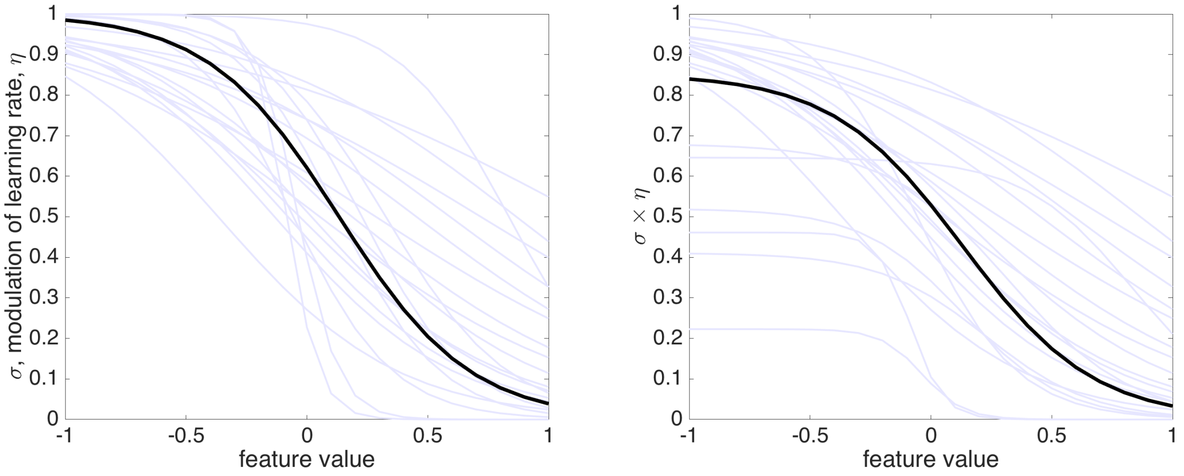
still describe subject choice better than or equivalent to the other four models. Figure 7 shows the relative feature values, $\Phi(f)$, for two example games (one learned and one unlearned game) for the subject with the highest β_2 value ($\beta_2 = 88.0299$, average likelihood per trial = 0.5574; next best likelihood per trial (fRL+Decay+DAS) = 0.5501) and the subject with the lowest β_2 value ($\beta_2 = 1.064$, average likelihood per trial = 0.5482; next best likelihood per trial (fRL+Decay+FA) = 0.5482). Subjects with high β_2 are more likely to exhibit greedy attention, quickly selecting, narrowing, and then discarding attention on a single feature. These subjects appear to have rapid fluctuations in attention. Meanwhile, subjects with lower β_2 parameters appear to have broader attention until sufficient evidence is collected for a single feature. Even then, the magnitude of relative feature attention is much lower for the low β_2 case.

When looking at the mean likelihood per trial of the fRL+Decay+FAS model as compared to the fRL+Decay, the 95% confidence interval of difference between fRL+Decay+FAS and fRL+Decay is [.0087,.0127], $p < 5e-10$, t-stat: 11.1690 (Figure 6b). This confirms that the new model indeed improved upon the likelihood per trial of the old model. It is plausible to think that fRL+Decay+FAS improved performance only during a short interval of the learning phase, but Figure 6c shows that for the comparison of trial-by-trial average likelihoods, fRL+Decay+FAS significantly outperformed fRL+Decay in all but the first four trials in a game.

A model which only gated learning and did not bias choice did not significantly outperform the fRL+Decay model (results not graphed, $p > .1$). All these results combined imply that, while biasing choice greatly aids the performance of the models, the combination of selective attention at choice and gating plasticity of learning by some function of the feature weight captures important characteristics of the brain's underlying mechanism.

Looking more closely at the fitted parameters of the sigmoid revealed that features with higher weight values have lower learning rates (Figure 8). All fitted slopes were negative, with a t-test showing that the group's mean slope was significantly below zero (one tailed t-test, $p < 1e-12$). This makes sense using the logic from the theoretical work which motivated this modulation (See Section 4.1.2). That is, the theory derived from [Roelfsema and van Ooyen, 2005] was that weights contributing more to the net prediction of reward should be more robust to noise. The models converged precisely to this behavior.

Curiously, small weights converged to high modulation of learning rates (Figure 8). While large weights (which likely contribute more to reward prediction) are gated against learning, lower weights are highly susceptible to prediction errors. Let us consider behavior with positive prediction errors. When all feature weights are low (e.g., at the start of a game), there is not much evidence as



(a)

(b)

Figure 8: Modulating Learning Rate in fRL+Decay+FAS. (a) The lighter lines are the fitted sigmoids, σ, of the 21 subjects used to modulate their learning rate in the fRL+Decay+FAS model. The dark line is the sigmoid from taking the average slope and average intercept across subjects. (b) The lighter lines are the fitted sigmoids, σ, of the 21 subjects used to modulate learning rate in the fRL+Decay+FAS model multiplied by the respective subject's fitted learning rate. This represents exactly how a prediction error affects the old weight value of a feature as a function of the feature's old weight. (See Section 4.1.2 for how the sigmoid modulates learning rate). In general, the trend shows that large weights are more robust to noisy prediction errors, and small weights are more susceptible to prediction errors.

to which feature is most rewarding. In this scenario, according to [Dayan et al., 2000], the learning rate should be in fact be high. If positive predictions errors (i.e. one feature grows in weight and according to winner-takes-all dynamics is highly contributing), the magnitude of the prediction error is small. Even though learning is high for small weights, they are not updated as rapidly as when all weights are small. This persistent learning of features with smaller weights allows for some record of previously rewarded choices if confidence in the current representation wanes. High learning rate does not mean that low weighted features will suddenly acquire large weights and surpass that of the highest weighted feature. Rather, this sigmoidal gating still preserves the general hierarchy of weights representing the current task space while still allowing learning of alternate hypotheses.

In the face of negative prediction errors, this modulation of learning rate also has statistical advantages. As previously mentioned, if a weight is large, it means there is strong evidence it is

Model Name	Parameters	Mean (SD)	Range
fRL+Decay	η (learning rate)	0.0972 ± 0.0237	0 to 1
	β (softmax temperature†)	11.086 ± 1.811	0 to ∞
	d (decay rate)	0.4929 ± 0.1498	0 to 1
fRL+Decay+DA	η (learning rate)	0.2854 ± 0.0914	0 to 1
	β (softmax temperature†)	9.3875 ± 5.2098	0 to ∞
	d (decay rate)	0.4180 ± 0.1684	0 to 1
	β_2 (softmax temperature*)	4.9007 ± 8.4394	0 to ∞
fRL+Decay+FA	η (learning rate)	0.5633 ± 0.3362	0 to 1
	β (softmax temperature†)	11.847 ± 16.552	0 to ∞
	d (decay rate)	0.3809 ± 0.1716	0 to 1
	β_2 (softmax temperature*)	7.6281 ± 11.8265	0 to ∞
fRL+Decay+DAS	η (learning rate)	0.6308 ± 0.3399	0 to 1
	β (softmax temperature†)	7.8424 ± 4.1452	0 to ∞
	d (decay rate)	0.3829 ± 0.1857	0 to 1
	β_2 (softmax temperature*)	29.2568 ± 113.4108	0 to ∞
	A (slope of σ)	-6.2498 ± 6.5935	-50 to 50
	B (intercept of σ)	-3.1866 ± 4.0748	-50 to 50
fRL+Decay+FAS	η (learning rate)	0.8524 ± 0.2494	0 to 1
	β (softmax temperature†)	7.5677 ± 3.6353	0 to ∞
	d (decay rate)	0.2578 ± 0.1663	0 to 1
	β_2 (softmax temperature*)	12.2973 ± 19.4231	0 to ∞
	A (slope of σ)	-3.6989 ± 3.3707	-50 to 50
	B (intercept of σ)	-0.4917 ± 1.0619	-50 to 50

Table 1: Parameter table of fitted learning models according to 21 subjects. †softmax inverse temperature for calculating choice probability; *softmax inverse temperature for biasing choice

highly rewarding. It is important to preserve this representation against environmental noise and lower the feature's learning rate. Meanwhile, high learning of low weights with negative prediction errors also is statistically sound. When all feature weights are low, this high learning allows for a speedy process of elimination in finding the most rewarding feature. When one feature in the chosen stimulus has a high weighter, this means a larger negative prediction error if reward is omitted. This creates a larger divide between the more highly weighted feature and the smaller weighted features. This may allow a subject to better perform hypothesis testing by filtering out distractors.

It is interesting to note that for the best fit parameters of the fRL+Decay+FAS model, 13 of the 21 subjects' learning rates, η, converged to one. Recall that a subject updates the weights of the chosen stimulus, where $\Delta W(f) = \sigma\big(W(f)\big)\eta(R - V(S))$ (see Equation 11). This means that this sigmoid function entirely drove learning in these subjects. How the σ modulates learning rate can be seen in difference in sigmoidal shapes between Figure 8a and 8b. Other analyses show that fixing the learning rate, η, of the $fRL + Decay + FAS$ to one significantly outperforms $fRL + Decay + FAS$ on the BIC metric (due to lower model complexity, $p < 5e-08$), while the likelihoods per trial are not significantly different than $fRL + Decay + FAS$ (analysis not pictured). In general, these results show that the weight of a feature in the Dimensions Task acts as a gating variable for the plasticity and sensitivity of that weight to further learning.

4.5. Predicting Memory Probe

To ensure our fRL+Decay+FAS winning model still maintains fRL+Decay's predictive power for probe accuracy, we fixed the parameters found for the above model comparison. We then used the model to predict probe accuracy. This was also done for the fRL+Decay model. Again, similar to the original behavioral analysis of probe behavior (Section 3.3.1), we built models to separately predict chosen and unchosen stimuli accuracy due to differences in encoding. We compared these models' performances to a null memory model described below. For a reminder of the memory probe, see Section 3.2.

4.5.1. Chosen Stimulus

For memory of the chosen stimulus, we predicted the probability a subject would correctly recall the feature of the chosen stimulus with a logistic sigmoid, in correspondence with the behavioral results in Figure 3. The logistic was a function of feature weight or attention. The general structure for the models is:

$$P(\text{Memory of } f \in S_{\text{chosen}}, D_{\text{probed}}) = \frac{1}{1 + e^{-A*(X(f))+B}} \tag{14}$$

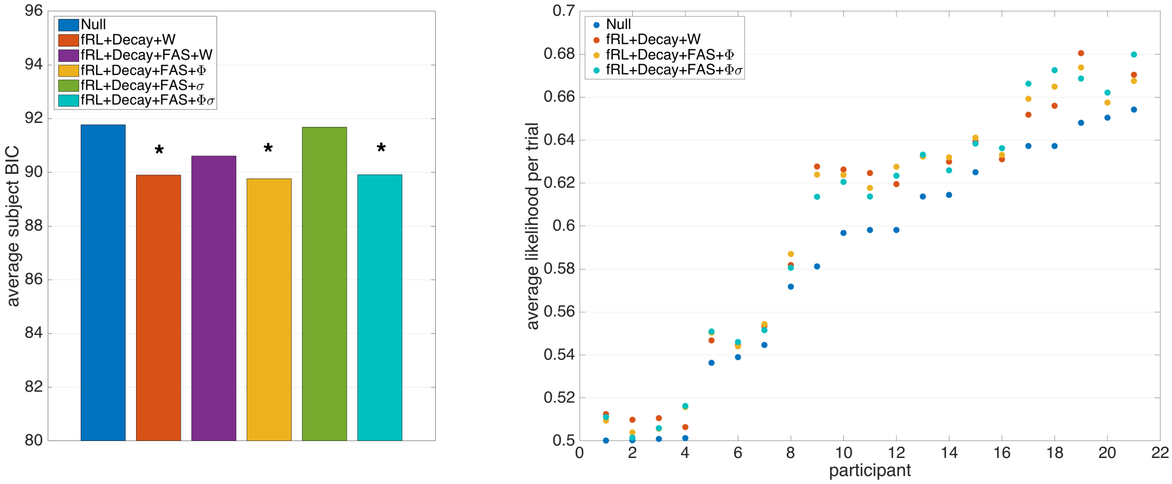
where $X(f)$ is some predictor related to the feature f. The models compared are as follows:

- *Null Memory Model* Our null model for the chosen assumed that participants would correctly recall the feature of the chosen stimulus when probed with some base probability, p, regardless of the feature's associative value or probability of predicting high reward. As we consider the null memory model independent of the learning and decision-making processes, the null model is the same for fRL+Decay and fRL+Decay+FAS.

- *fRL+Decay+W* where predictors are $X(f) = W(f)$ as calculated by the fRL+Decay model after feedback on the probed trial.
- *fRL+Decay+FAS+W* where predictors are $X(f) = W(f)$ as calculated by the fRL+Decay+FAS model after feedback on the probed trial.
- *fRL+Decay+FAS+*$\Phi(f)$ where predictors are $X(f) = \Phi(f)$ as calculated by the fRL+Decay+FAS model for choice on the probed trial. See Equation 8 for how relative attention of the feature, $\Phi(f)$, is calculated.
- *fRL+Decay+FAS+*$\sigma(f)$ where predictors are $X(f) = \sigma(f)$ as calculated by the fRL+Decay+FAS model for learning on the probed trial. See Equation 12 for calculation of gated learning, $\sigma(f)$, and Figure 8 for subject-level plots.
- *fRL+Decay+FAS+* $\Phi\sigma(f)$ where predictors are $X(f) = \Phi \times \sigma(f)$ as calculated by the fRL+Decay+FAS model for choice and learning on the probed trial.

As shown by Figure 9, while the fRL+Decay+FAS+W did not significantly outperform the null memory model, the fRL+Decay+FAS with attention metrics as predictors did. fRL+Decay+FAS+Φ ($p < 0.01$) and fRL+Decay+FAS+Φσ ($p < 0.005$), significantly outperformed the null model. The performances of fRL+Decay+FAS+Φ and fRL+Decay+FAS+Φσ were not significantly different than fRL+Decay+W. (fRL+Decay+W was significantly different than the null, $p < 0.05$); it is probable that the weights in fRL+Decay capture some of the probability information contained in Φ. The positive logistic slopes (Table 2) - i.e. larger attention metric, higher memory prediction - further support the theory which argues attention is allocated to the more predictive stimuli [Mackintosh, 1975, Dayan et al., 2000].

To extend this further, these results show that more predictive stimuli are not only more attended but better *encoded* in the brain, evidence of attention's role in constructing a simplified representa-



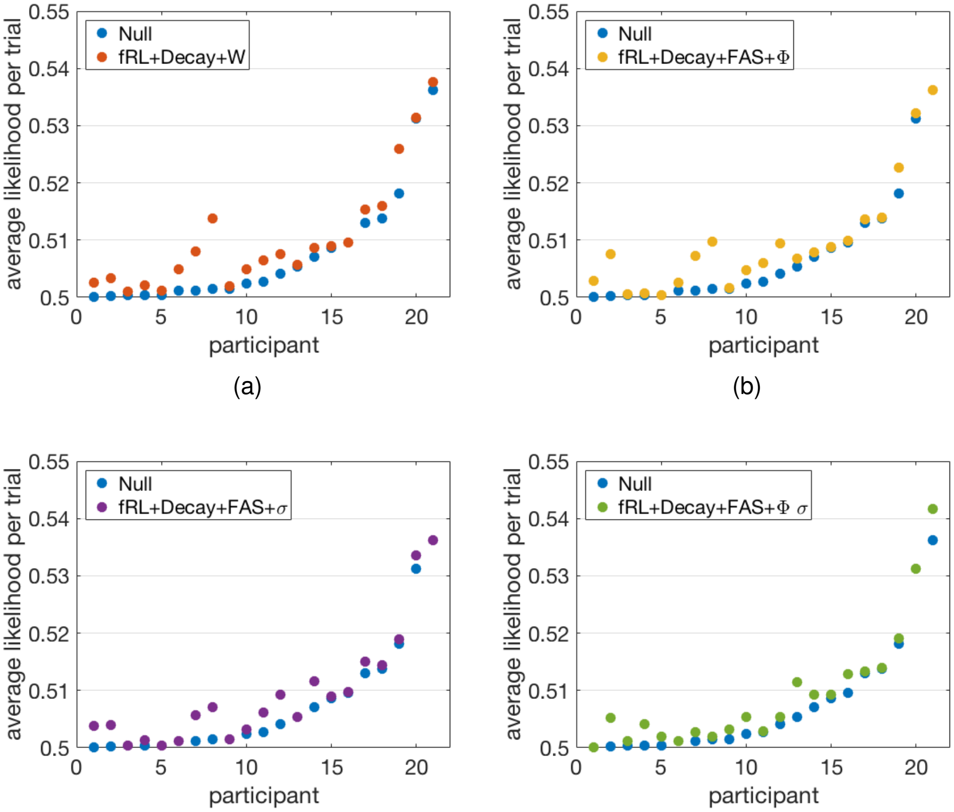
(a)

(b)

Figure 9: Model Comparison for Predicting Probe Accuracy of the Chosen Stimulus. (a) BIC comparison. * indicates the BIC of the model is significantly less than that of the null. The weights of the fRL+Decay model continued to be good predictors of probe accuracy on the chosen stimulus. They were the weights used in the behavioral results expressed in Figure 3. The weights of the fRL+Decay+FAS models, however, were not sufficient to outperform the null model. Instead, the attentional metric at choice, Φ, and Φ times the attention at learning, $\Phi \times \sigma$, served as sufficient predictors of memory. Performance for the three starred models were not significantly different. (b) A plot of the average likelihood per trial for the three models that outperformed the null model.

Model Name	Parameters	Mean (SD)	Range
Null	p (baseline memory)	0.7370 ± 0.1235	0 to 1
fRL+Decay+W	A (slope of sigmoid)	5.5765 ± 2.8795	-50 to 50
	B (intercept of sigmoid)	-0.6139 ± 0.5461	-50 to 50
fRL+Decay+FAS+W	A (slope of sigmoid)	1.8704 ± 1.5975	-50 to 50
	B (intercept of sigmoid)	-0.6525 ± 0.6115	-50 to 50
fRL+Decay+FAS+Φ	A (slope of sigmoid)	3.5039 ± 2.0763	-50 to 50
	B (intercept of sigmoid)	-0.6170 ± 0.5284	-50 to 50
fRL+Decay+FAS+σ	A (slope of sigmoid)	-2.9096 ± 1.7168	-50 to 50
	B (intercept of sigmoid)	-2.3600 ± 1.1913	-50 to 50
fRL+Decay+FAS+$\Phi\sigma$	A (slope of sigmoid)	17.1653 ± 8.4592	-50 to 50
	B (intercept of sigmoid)	-0.4291 ± 0.5685	-50 to 50

Table 2: Parameter table of fitted memory models for chosen stimulus according to 21 subjects.



(a)

(b)

(c)

(d)

Figure 10: Predicting memory of unchosen stimuli of dimension probe using (a) weight predictors (fRL+Decay) or (b-d) attentional predictors (fRL+Decay+FAS). While likelihoods of the four graphs are significantly above the null (ps < 5e-04), BICs are significantly higher, meaning the models did not outperform the null model (ps < 1e-10).

tion of a complex environment. It is important to note that σ alone was insufficient for predicting memory; however, it did not tarnish performance (and for some subjects helped predict memory) when combined with Φ. This suggests that while encoding and memory may be driven largely by how predictable the feature is of high reward, that metric alone may not solely dictate representation of the trial.

4.5.2. Unchosen Stimulus Predicting accuracy on the unchosen stimuli is less clear. Weights of fRL+Decay in the original behavioral analysis were not predictive of the accuracy of unchosen stimuli. To simplify this analysis, we looked at predicting the probability that both unchosen stimuli were correct. As this is dependent on the chosen stimuli being correctly placed in the memory probe, we fit the data on all tested models only over those trials in which the chosen stimulus was correctly placed. The null model again assumed that the unchosen stimuli conditioned on the chosen stimulus being correct would also be placed correctly with some baseline probability, p. No model variant of either $fRL+Decay$ or $fRL+Decay+FAS$ that was tested successfully outperformed this null model. The weight or attention metrics of the feature in the chosen stimulus of the probed dimension were used as predictors in the same way as when predicting chosen probe accuracy. While the likelihood for some models was significantly higher than the null as seen in Figure 10, the likelihood payoff is not enough to offset the added complexity of the model. Any effect on attention of the unchosen stimulus according to feature learning appears to be minimal and nuanced, though some trending effects seem to emerge on a subject-by-subject basis.

5. Discussion

The work in this paper offers an improved computational model of the Dimensions Task, $fRL+Decay+FAS$, supported by model comparison evidence. The Dimensions Task, a multi-armed bandit task, is used to simulate learning in a noisy, high dimensional environment. The new model incorporates statistical ideas of attention derived from behavioral findings (some of which are presented in this paper) and grounded in theoretical thinking. Specifically, the model combines the concept of attention competitively allocated to stimuli according to their relative reliability of predicting reward [Mackintosh, 1975, Dayan et al., 2000], and gated learning of features [Roelfsema and van Ooyen, 2005, Dayan et al., 2000, Pearce and Hall, 1980]. Specifically, it does this by fusing the concepts presented in Dayan et al. (2000) - which posits that competitive predictability between stimuli biases calculations of compound expected reward - with softmax winner-takes-all dynamics and robustness of high contributing features to feedback from Roelfsema and van Ooyen (2005). By fitting the model to subject data and outperforming the fRL+Decay model, this work offers empirical evidence of these theoretical concepts and supports the integrated, bidirectional relationship of attention and learning proposed in [Leong et al., 2017]. Notably, this reinforcement learning model relies solely on the current state according to RL values, rather than

Bayesian inference, for selecting subsets of the feature space to learn and to attend.. Despite basic reinforcement learning's poor scaling to high dimensions, introducing attentional processing aids in improving performance and plausibility.

Behaviorally, this work shows that, not only do subjects attend more highly to more predictive features of a stimulus as other empirical work has shown, highly predictive features are *encoded* more strongly in the memory of the stimulus. This novel result further supports the implication of attention in discerning which features of a multidimensional space to include in an internal representation. An extension of this work would be to confirm if this model could predict a more rigorous metric of attention, such as eye-tracking or MVPA analysis, rather than memory. While this study shows that attention measures in fRL+Decay+FAS can predict memory, as discussed in Section 2.2.1, memory is an imperfect measure of attention.

To better understand how stimuli presented at choice are encoded, we propose two possible variants of the memory probe. The first design uses the same stimulus and probe design, but rather than probing after feedback, subjects should be probed directly after choice. Feedback should be omitted for the probed trial. This approach will better disentangle how a multidimensional space is encoded as the result of top-down processing. To complement this design, a separate group of participants should be given the paradigm described in this study, but both unchosen stimuli should remain on screen during feedback. This will allow for both chosen and unchosen stimuli to be presented for the same time duration. This will possibly yield a better metric of how unchosen stimuli are encoded during learning and may produce a clearer relationship between reinforcement learning values and encoding of irrelevant stimuli. This dual-paradigm will better differentiate encoding at choice versus learning.

Another variant to the memory probe is to ask subjects to report the three features of a specific stimulus. This variant can better address how individual features of the same stimulus are preferentially processed. Notably, unlike the dimensional probe presented in this paper, performance of recall of the three probed features is independent, in that color location information does not inform frequency location information. The σ of the fRL+Decay predicts updating differences in the lower weighted features of the chosen stimulus according to the magnitude of the prediction error. Specifically, if one feature in the chosen stimulus is high and the others are lower, in the event of no reward (i.e. large negative prediction error), the model would predict higher learning from prediction error for the lower-valued features. This in turn might correspond to amplified attention compared to when the prediction error is small. This hypothesis is in line with work by [Rouhani et al., 2017], which shows larger prediction errors lead to stronger episodic memory. The current design is not sufficient for this analysis.

Another interesting variation of the Dimensions Task to test extrapolated performance of the model and its attention predictions would be a gradient of different reward probabilities and a design where subjects are not told how reward is determined. For different reward probabilities, we hypothesize that less noise in reward (e.g. if selecting the highly rewarding feature was rewarded 80% of the time, else reward is delivered 10% of the time) will lead to larger β_2 values and flatter logistic slopes. That is, subjects will exhibit greedier behavior in calculating probabilities and be more sensitive to small changes in weight values. Subjects will also update lower weights less due to lower noise in the environment.

If subjects are not told how reward is determined in a game and reward structure is left ambiguous (e.g. if a stimuli contains any one of a subset of features, the stimulus is highly rewarding), we hypothesize this model, as it is, will not perform well. This is because the model takes advantage of the fact that the reward reliabilities and certainties of features are dependent - only one feature is highly rewarding. The fRL+Decay+FAS model exploits information provided to the subject, and in the absence of this assumption, a more general model is needed. (However, it is not unreasonable that subjects adapt learning styles based on assumptions about the structure of the environment.) The attentional framework proposed in the model can be extended, though how probabilities for calculating expected value of a compound stimulus would need to be modified. Given the difficulty subjects have in the original Dimensions Task, it is probable the noise in feedback will need to be lessened for this variation.

It would be exciting to extend the gated learning of the model with some calculation of feature uncertainty. This could be done with a running calculation of the average prediction error for a feature weighted by recency. The current model does not truly capture the uncertainty attention expressed in the Pearce-Hall model and Dayan et al. (2000) model, except whatever uncertainty measure may be captured by feature weights.

Finally, an ignored aspect of selective attention in this model (that was briefly hinted at in Section 2) is selective attention due to low-level salience. As stated in the behavioral results section of Section 2, low-level salience seems to play a large role in encoding if the dimension (e.g., color) is highly salient. While the focus of this paper was to understand how top-down attention due to associative learning affects stimulus encoding, a more unified model would integrate all three levels of selective attention - exogenous, endogenous, and value-driven attention. Specifically, it would be interesting to model how highly salient, though irrelevant features, change the magnitude of associative learning's influence on encoding of multi-dimensional stimuli.

The implications of this research fall into the broader view of representation learning and the brain. It further implicates the ongoing interaction between attention and learning as the brain's mechanism for countering the curse of dimensionality. It suggests key characteristics of this mechanism include statistical competition to encode an environment and gated updating of this encoding process modulated by confidence in state representation. The bi-directionality of these characteristics gives us the ability to filter the world according to current beliefs of what will most inform success in our goals, while allowing us sensitivity to environmental changes and adaptability if confidence in these internal beliefs shifts. This type of cyclical dynamic is essential for learning effectively in a complex, noisy, and ever-changing world.

6. Appendix

A. Original Dimensions Task

Reproduced below are example stimuli from the original Dimensions Task design by Niv et al. 2015 [Niv et al., 2015]. See paper for further details of study.

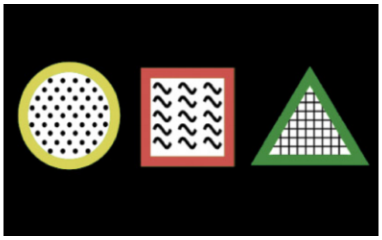
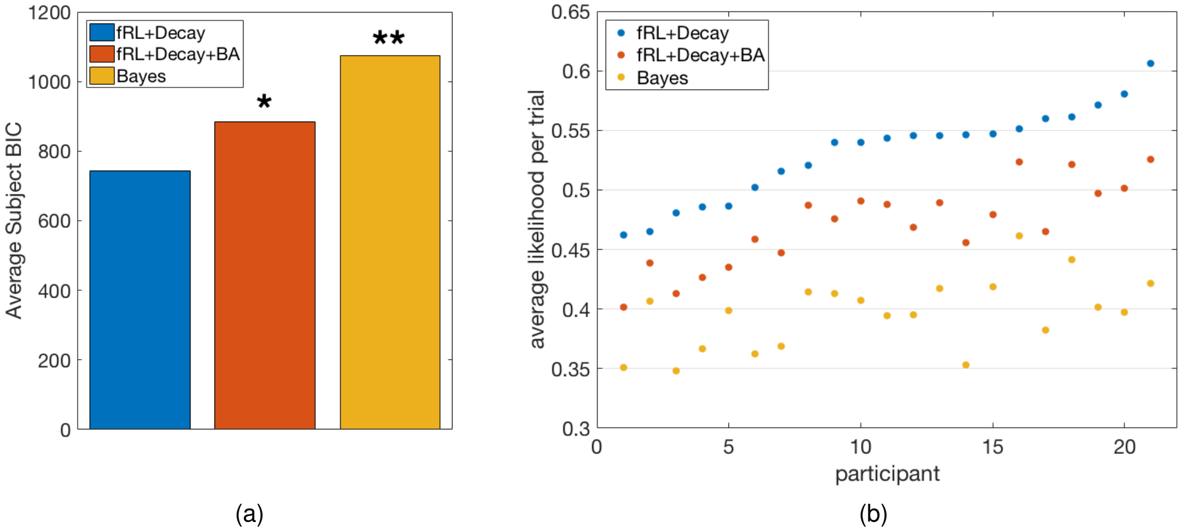


Figure 11: Example of original Dimensions Task stimuli

B. Modeling with Bayes' Theorem

This paper focused exclusively on modeling the Dimensions Task with reinforcement learning algorithms. An alternative way to model this task is by using Bayes' Theorem to track the likeli-



(a)

(b)

Figure 12: Model Comparison with Bayesian Models. Despite using a statistically optimal calculation of the probability a feature is highly rewarding, the fRL+Decay+BA continues to under perform in comparison to the fRL+Decay model. The Bayes model, in confirmation of the findings in [Niv et al., 2015] and [Geana and Niv, 2014], under performs both the fRL+Decay and fRL+Decay+BA models.

hood of a feature being the most highly rewarding. As evidence points to in [Niv et al., 2015] and [Geana and Niv, 2014], the brain likely does not implement Bayesian methods for representation learning. However, as a source of reference, in this section we present a model comparison between:

- *fRL+Decay*
- *Bayes* A Bayesian model which updates the probability of the feature being the most highly rewarding feature, f^*, rather than calculating feature weights.
- *fRL+Decay+BA* Instead of using a softmax to calculate the probability of a feature being highly rewarding to bias choice (Equation 7), this model instead uses the probabilities in the *Bayes* model above. This is similar to the Hybrid Bayes-RL model in [Niv et al., 2015], though does not update learning.

For calculation of Bayesian probabilities, see [Geana and Niv, 2014].

As seen in Figure 12, the fRL+Decay model continues to outperform both of the Bayesian models (BIC comparison: (fRL+Decay+BA) $p < 1e-11$, (Bayes) $p < 5e-13$). Of particular note is that while fRL+Decay+BA improves upon the Bayes model ($p < 1e-12$), it remains significantly outperformed by the fRL+Decay model. This is in direct contrast to when values of stimuli are biased by probabilities calculated by a softmax function of all feature weights, which improved fRL+Decay performance.

C. Honor Code

This paper represents my own work in accordance with University regulations.

/s/ Alana Jaskir

References

[Anderson, 2015] Anderson, B. A. (2015). The attention habit: How reward learning shapes attentional selection. *Annals of the New York Academy of Sciences*.

[Anderson et al., 2011] Anderson, B. A., Laurent, P. A., and Yantis, S. (2011). Value-driven attentional capture. *Proceedings of the National Academy of Sciences*, 108(25):10367–10371.

[Beierholm et al., 2009] Beierholm, U. R., Quartz, S. R., and Shams, L. (2009). Bayesian priors are encoded independently from likelihoods in human multisensory perception. *Journal of vision*, 9(5):23–23.

[Bellman, 1957] Bellman, R. (1957). *Dynamic Programming*. Princeton University Press.

[Bengio et al., 2013] Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828.

[Bengio et al., 2015] Bengio, Y., Lee, D.-H., Bornschein, J., Mesnard, T., and Lin, Z. (2015). Towards biologically plausible deep learning. *arXiv preprint arXiv:1502.04156*.

[Bucker et al., 2015] Bucker, B., Belopolsky, A. V., and Theeuwes, J. (2015). Distractors that signal reward attract the eyes. *Visual Cognition*, 23(1-2):1–24.

[Buschman and Kastner, 2015] Buschman, T. J. and Kastner, S. (2015). From behavior to neural dynamics: an integrated theory of attention. *Neuron*, 88(1):127–144.

[Cohen et al., 2007] Cohen, J. D., McClure, S. M., and Angela, J. Y. (2007). Should i stay or should i go? how the human brain manages the trade-off between exploitation and exploration. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 362(1481):933–942.

[Dayan et al., 2000] Dayan, P., Kakade, S., and Montague, P. R. (2000). Learning and selective attention. *nature neuroscience*, 3:1218–1223.

[Geana and Niv, 2014] Geana, A. and Niv, Y. (2014). Causal model comparison shows that human representation learning is not bayesian. In *Cold Spring Harbor symposia on quantitative biology*, volume 79, pages 161–168. Cold Spring Harbor Laboratory Press.

[Gershman and Niv, 2010] Gershman, S. J. and Niv, Y. (2010). Learning latent structure: carving nature at its joints. *Current opinion in neurobiology*, 20(2):251–256.

[Jones and Canas, 2010] Jones, M. and Canas, F. (2010). Integrating reinforcement learning with models of representation learning. In *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*, pages 1258–1263.

[Körding and Wolpert, 2004] Körding, K. P. and Wolpert, D. M. (2004). Bayesian integration in sensorimotor learning. *Nature*, 427(6971):244–247.

[Le Pelley et al., 2015] Le Pelley, M. E., Pearson, D., Griffiths, O., and Beesley, T. (2015). When goals conflict with values: Counterproductive attentional and oculomotor capture by reward-related stimuli. *Journal of Experimental Psychology: General*, 144(1):158.

[Lee and Shomstein, 2014] Lee, J. and Shomstein, S. (2014). Reward-based transfer from bottom-up to top-down search tasks. *Psychological Science*, 25(2):466–475.

[Leong et al., 2017] Leong, Y. C., Radulescu, A., Daniel, R., DeWoskin, V., and Niv, Y. (2017). Dynamic interaction between reinforcement learning and attention in multidimensional environments. *Neuron*, 93(2):451–463.

[Mackintosh, 1975] Mackintosh, N. J. (1975). A theory of attention: Variations in the associability of stimuli with reinforcement. *Psychological review*, 82(4):276–298.

[Marković et al., 2015] Marković, D., Gläscher, J., Bossaerts, P., O'Doherty, J., and Kiebel, S. J. (2015). Modeling the evolution of beliefs using an attentional focus mechanism. *PLoS Comput Biol*, 11(10):e1004558.

[McMains and Kastner, 2011] McMains, S. and Kastner, S. (2011). Interactions of top-down and bottom-up mechanisms in human visual cortex. *Journal of Neuroscience*, 31(2):587–597.

[Mine and Saiki, 2015] Mine, C. and Saiki, J. (2015). Task-irrelevant stimulus-reward association induces value-driven attentional capture. *Attention, Perception, & Psychophysics*, 77(6):1896–1907.

[Niv et al., 2015] Niv, Y., Daniel, R., Geana, A., Gershman, S. J., Leong, Y. C., Radulescu, A., and Wilson, R. C. (2015). Reinforcement learning in multidimensional environments relies on attention mechanisms. *The Journal of Neuroscience*, 35:8145—-8157.

[Pearce and Hall, 1980] Pearce, J. M. and Hall, G. (1980). A model for pavlovian learning: variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological review*, 87(6):532.

[Pearson et al., 2015] Pearson, D., Donkin, C., Tran, S. C., Most, S. B., and Le Pelley, M. E. (2015). Cognitive control and counterproductive oculomotor capture by reward-related stimuli. *Visual Cognition*, 23(1-2):41–66.

[Roelfsema and van Ooyen, 2005] Roelfsema, P. R. and van Ooyen, A. (2005). Attention-gated reinforcement learning of internal representations for classification. *Neural computation*, 17(10):2176–2214.

[Rouhani et al., 2017] Rouhani, N., Norman, K. A., and Niv, Y. (2017). Dissociable effects of surprising rewards on learning and memory. *bioRxiv*, page 111070.

[Sali et al., 2014] Sali, A. W., Anderson, B. A., and Yantis, S. (2014). The role of reward prediction in the control of attention. *Journal of experimental psychology: human perception and performance*, 40(4):1654.

[Schuck et al., 2016] Schuck, N. W., Cai, M. B., Wilson, R. C., and Niv, Y. (2016). Human orbitofrontal cortex represents a cognitive map of state space. *Neuron*, 91(6):1402–1412.

[Schwarz et al., 1978] Schwarz, G. et al. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464.

[Shin and Ma, 2016] Shin, H. and Ma, W. J. (2016). Crowdsourced single-trial probes of visual working memory for irrelevant features. *Journal of vision*, 16(5):10–10.

[Sutton and Barto, 1998] Sutton, R. S. and Barto, A. G. (1998). *Introduction to Reinforcement Learning*. MIT Press.

[Vaidya and Fellows, 2016] Vaidya, A. R. and Fellows, L. K. (2016). Necessary contributions of human frontal lobe subregions to reward learning in a dynamic, multidimensional environment. *Journal of Neuroscience*, 36(38):9843–9858.

[Wilson and Niv, 2012] Wilson, R. C. and Niv, Y. (2012). Inferring relevance in a changing world. *Frontiers in human neuroscience*, 5:189.

[Wilson et al., 2014] Wilson, R. C., Takahashi, Y. K., Schoenbaum, G., and Niv, Y. (2014). Orbitofrontal cortex as a cognitive map of task space. *Neuron*, 81(2):267–279.