

# Reinforcement Learning in Multidimensional Environments Relies on Attention Mechanisms

Yael Niv,<sup>1</sup> Reka Daniel,<sup>1</sup> Andra Geana,<sup>1</sup> Samuel J. Gershman,<sup>2</sup> Yuan Chang Leong,<sup>3</sup> Angela Radulescu,<sup>1</sup> and Robert C. Wilson<sup>4</sup>

<sup>1</sup>Department of Psychology and Neuroscience Institute, Princeton University, Princeton, New Jersey 08540, <sup>2</sup>Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, <sup>3</sup>Department of Psychology, Stanford University, Stanford, California 94305, and <sup>4</sup>Department of Psychology and Cognitive Science Program, University of Arizona, Tucson, Arizona 85721

In recent years, ideas from the computational field of reinforcement learning have revolutionized the study of learning in the brain, famously providing new, precise theories of how dopamine affects learning in the basal ganglia. However, reinforcement learning algorithms are notorious for not scaling well to multidimensional environments, as is required for real-world learning. We hypothesized that the brain naturally reduces the dimensionality of real-world problems to only those dimensions that are relevant to predicting reward, and conducted an experiment to assess by what algorithms and with what neural mechanisms this “representation learning” process is realized in humans. Our results suggest that a bilateral attentional control network comprising the intraparietal sulcus, precuneus, and dorsolateral prefrontal cortex is involved in selecting what dimensions are relevant to the task at hand, effectively updating the task representation through trial and error. In this way, cortical attention mechanisms interact with learning in the basal ganglia to solve the “curse of dimensionality” in reinforcement learning.

**Key words:** attention; fMRI; frontoparietal network; model comparison; reinforcement learning; representation learning

## Introduction

To make correct decisions, we must learn from past experiences. Learning has long been conceptualized as the formation of associations among stimuli, actions, and outcomes—associations that can then guide decision making in the presence of similar stimuli. But how should stimuli be defined in complex, multidimensional, real-world environments? Naïvely, it would seem optimal to learn about all available stimuli, for example, all observable objects as defined by their features (e.g., height, color, shape). However, it is often the case that only a few dimensions are relevant to the performance of any given task. Imagine standing on a street corner: if your task is to cross the street, you will likely ignore the colors of the cars and concentrate on their speed and distance; however, if your task is to hail a taxi, you should take color into account and can ignore other aspects. Learning and basing decisions on only those dimensions that are relevant to the task at hand improves performance, speeds learning, and simplifies generalization to future situations.

The computational framework of reinforcement learning (RL) has had a tremendous impact on our understanding of the neural basis of trial-and-error learning and decision making. Most notably, it offers a principled theory of how basal-ganglia structures support decision making by learning the future reward value of stimuli (or “states” in RL terminology) using prediction errors that are conveyed by midbrain dopaminergic neurons (Barto, 1995; Montague et al., 1996; Schultz et al., 1997; Niv and Schoenbaum, 2008; Niv, 2009). However, the bulk of this work has concentrated on learning about simple stimuli. When stimuli are multidimensional, RL algorithms famously suffer from the “curse of dimensionality,” becoming less efficient as the dimensionality of the environment increases (Sutton and Barto, 1998). One solution is to select a small subset of dimensions to learn about. This process has been termed “representation learning” as it is tantamount to selecting a suitable state representation for the task (Gershman and Niv, 2010; Wilson and Niv, 2011).

Neurally, it is plausible that corticostriatal projections are shaped so as to include only stimulus dimensions that are presumed to be relevant to the task at hand (Bar-Gad et al., 2000), for instance by selective attention mechanisms (Corbetta and Shulman, 2002). Striatal circuits can also contribute to highlighting some inputs and not others (Frank and Badre, 2012). Such attentional filters, in turn, should be dynamically adjusted according to the outcomes of ongoing decisions (Cañas and Jones, 2010; Frank and Badre, 2012), forming a bidirectional interaction between representation learning and RL.

To study the neural basis of representation learning, we designed a “dimensions task”—a multidimensional bandit task in which only one of three dimensions (color, shape, or texture)

Received July 20, 2014; revised March 20, 2015; accepted March 27, 2015.

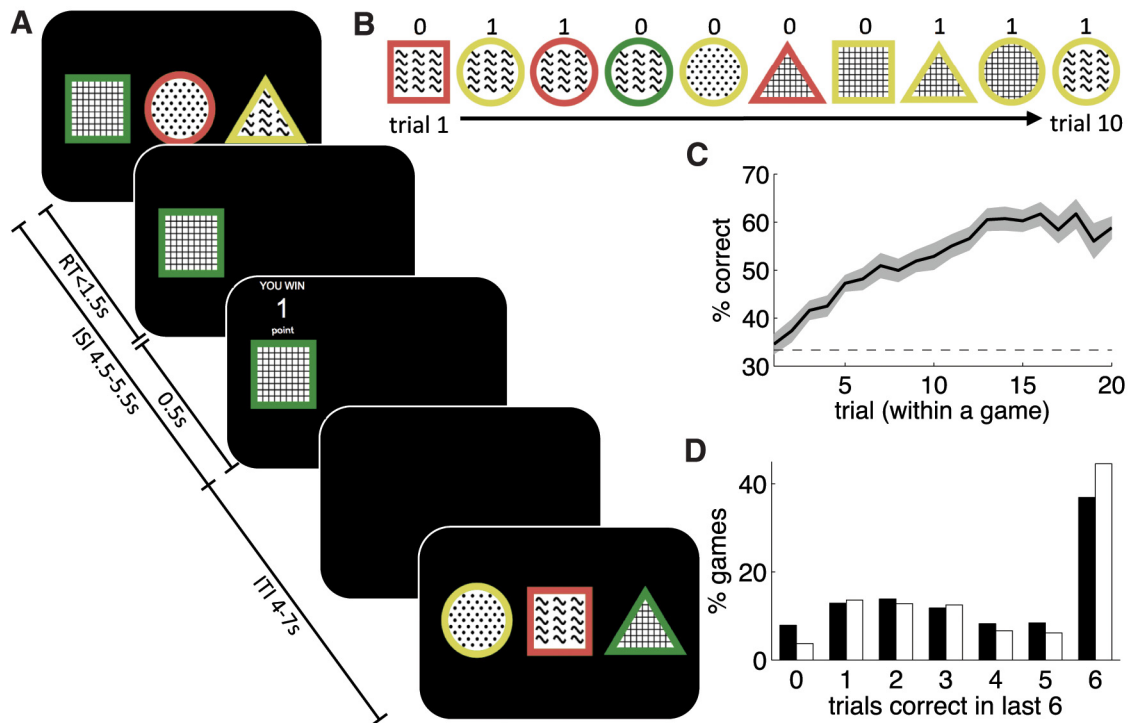
Author contributions: Y.N. designed research; Y.N. performed research; Y.N., R.D., A.G., S.J.G., Y.C.L., A.R., and R.C.W. analyzed data; Y.N., R.D., A.G., S.J.G., Y.C.L., A.R., and R.C.W. wrote the paper.

This research was supported by Award R03DA029073 from the National Institute of Drug Abuse, Award R01MH098861 from the National Institute of Mental Health, an Ellison Medical Foundation Research Scholarship to Y.N., and a Sloan Research Fellowship to Y.N. The content is solely the responsibility of the authors and does not represent the official views of the National Institutes of Health, the Ellison Medical Foundation, or the Sloan Foundation. We thank Michael T. Todd for helpful discussions in the early phases of the project; and Russell Poldrack and Todd Gureckis for comments on an earlier version of this manuscript.

Correspondence should be addressed to Yael Niv, Department of Psychology and Neuroscience Institute, Princeton University, Princeton, NJ 08540. E-mail: yael@princeton.edu.

DOI:10.1523/JNEUROSCI.2978-14.2015

Copyright © 2015 the authors 0270-6474/15/358145-13\$15.00/0



**Figure 1.** Task and behavioral results. **A**, Schematic of the dimensions task. Participants were presented with three different stimuli, each having a different feature along each one of the three dimensions (shape, color, and texture). Participants then selected one of the stimuli and received binary reward feedback, winning 1 (depicted) or 0 points. After a short delay, a new trial began with three new stimuli. **B**, Illustration of one game for one participant. Only the chosen stimulus is depicted for each of 10 consecutive trials, along with the outcome of each choice. **C**, Learning across games and participants, for games in the first 500 trials. Plotted is the percentage of choices of the stimulus that contained the target feature, throughout the games. Dashed line, chance performance; shaded area, SEM across participants. Learning in the 300 trials during functional imaging was similar, but the learning curve is less interpretable as games were truncated when a performance criterion was reached (see Materials and Methods). Other measures of learning, such as the number of trials to criterion (mean = 17.00 for the 500 fast-paced trials; mean = 16.40 for the slower-paced 300 trials;  $p = 0.09$ , paired  $t$  test), also suggest that performance in the two phases of the task was comparable. **D**, Percentage of games in which the stimulus containing the target feature was chosen on 0–6 of the last 6 trials of each game, across participants and games in the first 500 fast-paced trials (black) and in the last 300 slower-paced trials (white). In ~40% of the games, participants consistently chose the stimulus that contained the correct feature (6 of 6 trials correct), evidencing that they had learned the identity of the target feature. In the rest of the games, their performance correct was at chance (on average, only two trials containing the target stimulus, consistent with the participant “playing” on an incorrect dimension and only selecting the stimulus containing the target feature by chance, that is, one-third of the time).

determined reward. We scanned the brains of human participants as they played this task, changing the reward-relevant dimension frequently and fitting participants’ trial-by-trial choice data to the predictions of different computational models of learning. We then used the best model to generate regressors corresponding to the dynamics of representation learning and to search for neural areas that may be involved in this process.

## Materials and Methods

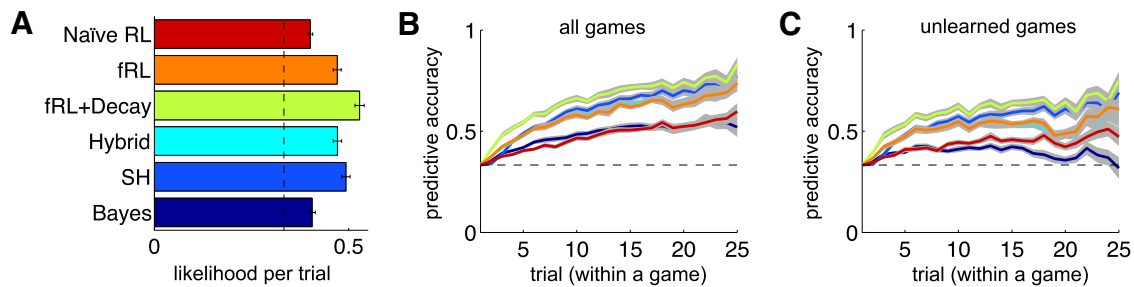
**Subjects.** Thirty-four participants (20 females; age range, 18–26 years; mean age, 20.9 years; 1 participant was left handed) were recruited from the Princeton University community. All participants gave informed consent and were compensated for their time at a rate of \$20/h (resulting in payments of \$35–50, with the majority of participants being paid \$40). Due to equipment failure, data for three participants were not complete and so were not analyzed. Another three participants who showed excess head movements (i.e., >4 mm in any direction) and six participants who failed to perform the task reliably above chance levels (i.e., missed  $\geq 50$  trials and/or had an overall performance of <38% correct during the functional scan, with the performance threshold calculated as the 95% confidence interval around random performance) were discarded from further analysis, leaving a total of 22 participants (15 females; age range, 18–26 years; mean age, 21.1 years; all participants were right handed) whose data are reported below. Study materials and procedures were approved by the Princeton University Institutional Review Board.

**Task.** Figure 1A shows a schematic of the task. On each trial, participants were presented with three stimuli, each consisting of one feature on each of the following three dimensions: shape (square, triangle, or circle);

color (red, green, or yellow); and texture (plaid, dots, or waves). The participant’s task was to choose one stimulus on each trial, with the goal of accumulating as many points as possible. After a stimulus was chosen, the other two stimuli disappeared from the screen, and, after a variable interstimulus interval (ISI), the outcome of the choice (either 1 or 0 points) appeared above the chosen stimulus. The outcome remained on screen for 500 ms, after which a fixation cross was displayed on a blank (black) screen for the duration of a variable intertrial interval (ITI).

Before performing the task, participants received on-screen instructions informing them that at any point only one of the three dimensions (color, shape, or texture) was relevant to determining the probability of winning a point, that one feature in the relevant dimension would result in rewards more often than the others (the exact probability was not mentioned), and that all rewards were probabilistic. Specifically, they were told that “One stimulus is better in that it gives you 1 most of the time (and 0 some of the time), the other two give you 0 most of the time (and 1 some of the time).” Participants were also instructed to respond quickly and to try to get as many points as possible. They first practiced three games in which the relevant dimension was instructed (one game in each dimension). They were then told that from now on they would not be instructed about the relevant dimension, and the actual task commenced.

The task consisted of a series of “games,” each 15–25 (uniform distribution) trials long. In each game, rewards were determined based only on one “relevant” dimension (color, shape, or texture). Within this dimension there was one “target” feature—selecting the stimulus that had this feature led to 1 point with 75% chance, and 0 points otherwise. The other two features in the reward-relevant dimension were associated with only a 25% chance of reward. The features in the two irrelevant dimensions



**Figure 2.** Model fits. **A**, Average likelihood per trial (when predicting the participant's choice on trial  $t$  given the choices and outcomes from the beginning of the game and up to trial  $t - 1$ ) for each of the six models. The model that explained the data best was the fRL + decay model. Error bars indicate the SEM. Dashed line, chance performance. **B**, Predictive accuracy (average likelihood per trial across games and participants) as a function of trial number within a game, for each of the models (colors are as in **A**; the hybrid model curve is almost completely obscured by that of the fRL model). By definition, all models start at chance. The fRL + decay model predicted participants' performance significantly better ( $p < 0.05$ ) than each of the other models from the second trial of the game and onward (excluding the 24th trial when comparing with fRL and SH, and the last two trials when comparing with hybrid), predicting participants' choices with  $>80\%$  accuracy by trial 25. **C**, These results hold even when considering only unlearned games, that is, games in which the participant chose the stimulus containing the target feature on fewer than 4 of the last 6 trials. Again, the predictions of the fRL + decay model were significantly better than those of the competing models from the second trial and onward (excluding the 24th trial when comparing with fRL and hybrid, and the 19th, 21st, and last two trials when comparing with hybrid). Moreover, the model predicted participants' behavior with  $>70\%$  accuracy by trial 25, despite the fact that participants' performance was not different from chance with respect to choosing the stimulus containing the target feature ( $p > 0.05$  for all but two trials throughout the game). The predictions of the Bayesian model, in contrast, were not statistically different from chance from trial 19 and onward, suggesting that this model did well in explaining participants' behavior largely due to the fact that both the model and participants learned the task. All data depicted are from the first 500 trials. Similar results were obtained when comparing models based on the 300 trials during the functional scan; however, the performance criterion applied in those games obscures the differences between learned and unlearned games, as seen in **B** and **C**, and thus those data are not depicted.

were inconsequential to determining reward. The relevant dimension in a particular game was always different from the relevant dimension in the previous game (participants were not told this).

For all data reported here, game transitions were signaled to the participant via a screen that declared that the previous game was over and a new game with a new relevant dimension is beginning. Fifteen of the 22 participants first performed 500 fast-paced trials (ISI of 0.3 s or the reaction time, the longer of the two; ITI of 0.3 s) outside the MRI scanner, and another 500 fast-paced trials during the 12 min structural scan. In these latter trials, game switches were not signaled. These unsignaled switch data have been reported elsewhere (Wilson and Niv, 2011) and are not analyzed further or reported here. The remaining seven participants performed the first 500 fast-paced trials with signaled dimension changes during the structural scan and did not perform the unsignaled switch task. In all cases, the structural scan was followed by four functional scans (see below), each consisting of 75 trials of the task with signaled dimension switches, with the timing of trials adjusted to the slow hemodynamic response [ISI, 4.5–5.5 s (inclusive of the reaction time, uniform distribution); ITI, 4–7 s (inclusive of outcome presentation time, uniform distribution)]. Throughout these 300 slower-paced trials, a criterion on performance was imposed such that after six consecutive choices of the stimulus with the target feature the game had a 50% likelihood of ending on any trial. Otherwise, games ended after a uniformly drawn length of 15–25 trials. Throughout, participants were required to make their choices within 1.5 s (3 s for the first participant). This response deadline was identical in the fast-paced and slower-paced phases of the experiment, and was imposed to invoke implicit decision making and RL processes rather than slow, deliberative processes. Failure to make a choice within the allotted time led to a time-out screen with the message “too slow,” and then to the ITI and the next trial (these missed trials were not counted toward the length of the game). Other variants of the task using longer response deadlines (up to 5 s) showed comparable results.

Our dimensions task combines elements from the Wisconsin Card Sorting Task (WCST), widely used to study cognitive flexibility (Milner, 1963; Lyvers and Maltzman, 1991; Owen et al., 1993; van Spaendonck et al., 1995; Ornstein et al., 2000; Lawrence et al., 2004; Monchi et al., 2004; Buchsbaum et al., 2005; Bishara et al., 2010), and the weather prediction task, a probabilistic categorization task that has been used to study and compare implicit and explicit learning processes (Gluck et al., 2002; Shohamy et al., 2004; Ashby and Maddox, 2005; Rodriguez et al., 2006; Price, 2009). By focusing on only one relevant dimension and frequent changes in that dimension, in combination with probabilistic rewards, we better emulated real-world learning (Kruschke, 2006),

prolonged the learning process so as to allow a detailed computational analysis of the dynamics of learning (see below), and ensured that prediction errors occur on every trial (Niv and Schoenbaum, 2008).

**Computational models.** To analyze learning dynamics in our task, we compared six qualitatively different computational models that reside on a spectrum ranging from ostensibly suboptimal to nearly statistically optimal solutions of the dimensions task (Fig. 2A). Below we describe each model. We begin with the less optimal of the models (“naïve RL”) and the most optimal of the models (“Bayesian learning”). These models will bracket and provide baseline comparisons for the four intermediate models that follow them.

**Naïve RL.** This model learns values for each of the 27 compound stimuli in the experiment using standard temporal difference (or Rescorla–Wagner) learning (Rescorla and Wagner, 1972; Sutton and Barto, 1998). Specifically, after choosing a stimulus,  $S$  (say, a green square with plaid texture), and observing the reward for that choice,  $R_t \in \{0,1\}$ , the value of that stimulus,  $V(S)$ , is updated according to the following:

$$V^{\text{new}}(S_{\text{chosen}}) = V^{\text{old}}(S_{\text{chosen}}) + \eta(R_t - V^{\text{old}}(S_{\text{chosen}})), \quad (1)$$

where  $\eta$  is a step-size or learning-rate parameter, and at the beginning of each game all 27 values are initialized at 0. To select one of the three stimuli available on each trial ( $S_1, S_2, S_3$ ), their current values are entered into a softmax probabilistic choice function, as follows:

$$p(\text{choose } S_i) = \frac{e^{\beta V(S_i)}}{\sum_{j=1}^3 e^{\beta V(S_j)}}, \quad (2)$$

such that the inverse temperature parameter  $\beta$  sets the level of noise in the decision process, with large  $\beta$  corresponding to low decision noise and near-deterministic choice of the highest-value option, and small  $\beta$  corresponding to high decision noise and nearly random decisions.

This model has two free parameters,  $\Theta = \{\eta, \beta\}$ , which we fit to each participant's data separately (see below). The model is naïve in the sense that it does not generalize between stimuli—what it learns about green squares with plaid texture does not affect the value of other green stimuli, at odds with the reward structure of the task. We thus used this model only as a baseline or null model for comparison with other, more sensible, strategies.

Note that here, and in all other reinforcement-learning based models we tested, values were initialized to 0. Given the instructions and experience with the task, one might assume that participants initialized their predictions at the start of each game with an intermediate value, between 0 and 1 point. However, treating the initial value as a free parameter and

fitting it to the data of each participant separately resulted in initial values that were close to 0 and no significant improvement of the performance of the models in predicting participants' behavior. The inverse Hessian of the likelihood of the data with respect to this parameter also suggested that the parameter was not well specified, in that large changes in initial value had very little effect on the overall likelihood of the data. Thus, we deemed the additional parameter not statistically justified, and used instead the more parsimonious version of each of the models, with initial values set (arbitrarily) at 0.

**Bayesian learning.** In contrast to the naïve RL model, the Bayesian learning model uses statistically optimal Bayesian inference, together with prior knowledge about the task as it was described to participants, to infer the probability of reward for each available stimulus. Specifically, the model tracks  $p(f = f^* | \mathcal{D}_{1:t-1})$ , the probability that each one of the nine features,  $f$ , is the target feature  $f^*$  given  $\mathcal{D}_{1:t-1} = \{C_{1:t-1}, R_{1:t-1}\}$ , the data (choices and rewards) from the beginning of the game and up to the current trial. The probability of each feature being the target feature is initialized at 1/9 at the beginning of a game and is subsequently updated after every trial according to Bayes' rule, as follows:

$$p(f = f^* | \mathcal{D}_{1:t}) \propto p(R_t | f = f^*, C_t) p(f = f^* | \mathcal{D}_{1:t-1}), \quad (3)$$

where the first argument on the right-hand side is  $p = 0.75$  or  $p = 0.25$ , depending on the reward on the current trial and whether the current choice  $C_t$  included  $f^*$ .

This probability distribution can be used to determine the value of choosing stimulus  $S$  as the probability of reward for choosing that stimulus on the current trial  $t$ :

$$V(S) = p(R_t = 1 | S, \mathcal{D}_{1:t-1}) = \sum_{f \in S} p(R_t = 1 | f = f^*, S) p(f = f^* | \mathcal{D}_{1:t-1}). \quad (4)$$

Here  $p(R_t = 1 | f = f^*, S) = 0.75$  for features  $f$  contained in  $S$ , and  $p(R_t = 1 | f = f^*, S) = 0.25$  for those that are not part of the evaluated stimulus. This model can be thought of as an "ideal observer" model as it maintains a full probability distribution over the identity of  $f^*$  and updates this distribution in a statistically optimal way. However, we note that for this model and for all others, we use a "softly ideal" action selection policy, the softmax policy described above. The model thus has only the softmax inverse temperature parameter as a free parameter,  $\Theta = \{\beta\}$ .

**Feature RL (fRL).** This model takes advantage of the fact that, in our task, features (not whole stimuli) determine reward, and uses reinforcement learning to learn the weights for each of the nine features. The value of stimulus  $S$  is calculated as the sum of the weights of its features,  $W(f)$ , as follows:

$$V(S) = \sum_{f \in S} W(f), \quad (5)$$

and the weights of the three features of a chosen stimulus are updated according to the following:

$$W^{\text{new}}(f) = W^{\text{old}}(f) + \eta [R_t - V(S_{\text{chosen}})] \quad \forall f \in S_{\text{chosen}}. \quad (6)$$

Feature weights are initialized at 0 at the beginning of each game (as mentioned above, fitting the initial value of weights did not improve fits to the data or change any of the results). As before, action selection proceeds via the softmax decision rule, and thus this model has two free parameters,  $\Theta = \{\eta, \beta\}$ .

**Feature RL with decay.** The models described above used all trials from the beginning of a game effectively and did not suffer from "forgetting" of any kind. However, this might not be the case in the human brain. To account for forgetting, we developed a feature RL with decay (fRL+decay) model that is identical to the fRL model described above, except that on every trial the weights of features that did not occur in the chosen stimulus are decayed to 0 with a rate,  $d$ , as follows:

$$W^{\text{new}}(f) = (1 - d)W^{\text{old}}(f) \quad \forall f \notin S_{\text{chosen}}. \quad (7)$$

The fRL+decay model thus has three free parameters,  $\Theta = \{\eta, d, \beta\}$ . (Here too, fitting two additional free parameters, the initial value of

feature weights at the beginning of games and the target for weight decay, did not change the results qualitatively.)

**Hybrid Bayesian-fRL model.** This model combines the fRL model with "dimensional attention weights" derived from the Bayesian model. That is, Bayesian inference (as described above) is used to track  $p(f = f^*)$ , the probability that each feature is the target feature. On each trial, these probabilities are summed across all features of a dimension and are raised to the power of  $\alpha$  to derive dimensional attention weights,  $\phi_d$ , for each of the dimensions:

$$\phi_d = \frac{1}{z} \left[ \sum_{f \in d} p(f = f^* | \mathcal{D}_{1:t-1}) \right]^\alpha, \quad (8)$$

where  $z$  normalizes  $\phi_d$  to sum up to 1. These dimensional weights are then used for weighing features when calculating the value of each stimulus, as follows:

$$V(S) = \sum_{d=1}^3 w(f_d) \phi_d, \quad (9)$$

with  $f_d$  being the feature in stimulus  $S$  in dimension  $d$ . Similarly, the updating of feature weights for the chosen stimulus is weighed by dimensional attention, as follows:

$$W^{\text{new}}(f_d) = W^{\text{old}}(f_d) + \eta (R - V(S_{\text{chosen}})) \phi_d \quad \forall f \in S_{\text{chosen}}. \quad (10)$$

Note, in contrast, that the fRL model weighs all features equally both in choice [i.e., in calculating  $V(S)$ ], and in learning. The hybrid model has three free parameters,  $\Theta = \{\eta, \alpha, \beta\}$ .

**Serial hypothesis (SH) model.** This final model, from Wilson and Niv (2011), has a different flavor from the above models. Here we assume that participants selectively attend to one feature at a time and, over the course of several trials, test the hypothesis that the attended feature is the correct feature. More concretely, when a participant chooses to attend to a certain feature on trial,  $n$ , we assume that from that trial on, until he decides to discard this hypothesis, he chooses the stimulus containing the candidate feature with probability  $1 - \varepsilon$  and chooses randomly otherwise. After every trial, the participant performs a Bayesian hypothesis test to determine whether to switch away from, or stick with, the current hypothesis, based on the reward history since choosing this feature as a candidate. This is done by computing the log ratio of the likelihood of the candidate feature being the target and the likelihood that it is not the target, as follows:

$$LR = \log \frac{p(f = f^* | R_{t-n+1:t})}{1 - p(f = f^* | R_{t-n+1:t})}, \quad (11)$$

where

$$p(f = f^* | R_{t-n+1:t}) \propto p(R_t | f = f^*) p(f = f^* | R_{t-n+1:t-1}), \quad (12)$$

which amounts to counting the rewards obtained on all trials in which the stimulus containing the candidate feature was selected since trial  $n$ , the trial in which the current hypothesis was selected. The log ratio is then entered into a softmax function to determine the probability of switching to a different (randomly chosen) target feature, as follows:

$$p(\text{switch}) = \frac{1}{1 + e^{\lambda(LR - \theta)}}. \quad (13)$$

The free parameters of this model are thus  $\Theta = \{\varepsilon, \lambda, \theta\}$ . While this model is simple to define, fitting its parameters is complicated by the fact that at any given time it is not straightforward to know what hypothesis (i.e., candidate feature) the participant is testing. However, it is possible to infer a distribution over the different hypotheses using Bayesian inference and an optimal change point detection algorithm (for details, see Wilson and Niv, 2011).

**Model fitting.** We used each participant's trial-by-trial choice behavior to fit the free parameters,  $\Theta_m$ , of each model,  $m$  (Table 1), and asked to what extent each of the models explains the participant's choices. Model

**Table 1. Free parameters for each of the models, and their best-fit values across the participant pool when fitting the first 500 fast-paced trials or the 300 slower-paced trials from the functional scans**

Model	Parameter	Mean (SD) first 500 trials	Mean (SD) last 300 trials	Range	Prior
Naïve RL	$\eta$ (learning rate)	0.431 $\pm$ 0.160	0.514 $\pm$ 0.231	0–1	None
	$\beta$ (softmax inverse temperature)	5.55 $\pm$ 2.30	4.85 $\pm$ 1.88	0– $\infty$	Gamma (2, 3)
Bayesian model	$\beta$ (softmax inverse temperature)	4.34 $\pm$ 1.13	5.15 $\pm$ 1.86	0– $\infty$	Gamma (2, 3)
fRL	$\eta$ (learning rate)	0.047 $\pm$ 0.029	0.076 $\pm$ 0.042	0–1	None
	$\beta$ (softmax inverse temperature)	14.73 $\pm$ 6.37	10.62 $\pm$ 4.92	0– $\infty$	Gamma (2, 3)
fRL + decay	$\eta$ (learning rate)	0.122 $\pm$ 0.033	0.151 $\pm$ 0.039	0–1	None
	$d$ (decay)	0.466 $\pm$ 0.094	0.420 $\pm$ 0.124	0–1	None
	$\beta$ (softmax inverse temperature)	10.33 $\pm$ 2.67	9.18 $\pm$ 2.16	0– $\infty$	Gamma (2, 3)
Hybrid	$\eta$ (learning rate)	0.398 $\pm$ 0.233	0.540 $\pm$ 0.279	0–1	None
	$\alpha$ ('steepness' of dimension weights)	0.340 $\pm$ 1.21	0.122 $\pm$ 0.129	0– $\infty$	None
	$\beta$ (softmax inverse temperature)	14.09 $\pm$ 6.96	11.84 $\pm$ 4.96	0– $\infty$	Gamma (2, 3)
SH	$\varepsilon$ (choice randomness)	0.071 $\pm$ 0.024	0.110 $\pm$ 0.039	0–1	None
	$\theta$ (sigmoid 'threshold')	−4.06 $\pm$ 1.72	−4.68 $\pm$ 1.91	−10 to 0	None
	$\lambda$ (sigmoid slope)	0.873 $\pm$ 0.412	0.732 $\pm$ 0.385	0– $\infty$	Gamma (2, 3)

Parameters fit to both phases of the experiment were similar; however, the performance criterion on games in the last 300 trials likely influenced parameters such as the softmax inverse temperature, as the proportion of trials in which participants could reliably exploit what they had learned was limited. Parameters were constrained to the ranges specified, and a Gamma distribution prior with shape 2 and scale 3 was used for the softmax inverse temperature in all models.

likelihoods were based on assigning probabilities to the choices of each participant on each of the  $T$  trials, as follows:

$$\mathcal{L} = p(C_{1:T} | \Theta_m) = \prod_{t=1}^T p(C_t | \mathcal{D}_{1:t-1}, \Theta_m). \quad (14)$$

Due to the differences in task parameters (ISI, ITI, and criterion) in the 500 prescan trials compared with the 300 functional scan trials, we fit the parameters of each model to each participant's prescan and functional-scan data separately. To facilitate model fitting, we used a regularizing prior that favored realistic values for the softmax inverse temperature parameter  $\beta$  and maximum a posteriori (rather than maximum likelihood) fitting (Daw, 2011). We optimized model parameters by minimizing the negative log posterior of the data given different settings of the model parameters using the Matlab function `fmincon`. Parameters fit to the functional scan trials were used to generate model-based regressors for fMRI analysis (see below), whereas parameters fit to the prescan trials were used to predict participants' behavior for the purposes of model comparison (see below). Table 1 summarizes the model parameters, their mean value (and SD) from the fit to data, and the range constraints and priors on each parameter.

**Model comparison.** To compare between the models based on their predictive accuracy, we used leave-one-game-out cross-validation on the 500 prescan trials (comparisons based on the functional scan trials gave similar results). In this method, for each participant, every model, and each game, the model was fit to the participant's choice data excluding that game. The model, together with the maximum a posteriori parameters, was then used to assign likelihood to the trials of the left-out game. This process was repeated for each game to obtain the total predictive likelihood of the data. We then calculated the average likelihood per trial for the model by dividing the total predictive likelihood by the number of valid trials for that participant. The likelihood per trial is an intuitive measure of how well the model predicts participants' choices, with a value of 1 indicating perfect prediction and 1/3 corresponding to chance performance. We used this quantity to compare between the models. Note that this cross-validation process avoids overfitting and allows direct comparison between models that have different numbers of parameters, as in our case.

**Imaging.** Brain images were acquired using a Siemens 3.0 tesla Allegra scanner. Gradient echo T2\*-weighted echoplanar images (EPIs) with blood oxygenation-level dependent (BOLD) contrast were acquired at an oblique orientation of 30° to the anterior–posterior commissure line, using a circular polarized head coil. Each volume comprised 41 axial slices. Volumes were collected in an interleaved-ascending manner, with the following imaging parameters: echo time, 30 ms; field of view, 191 mm; in-plane resolution and slice thickness, 3 mm; repetition time, 2.4 s. EPI data were acquired during four runs of 75 trials each and variable length. Whole-brain high-resolution T1-weighted structural scans (1 ×

1 × 1 mm) were also acquired for all participants and were coregistered with their mean EPI images. We note that with these imaging parameters, due to sometimes partial coverage of the whole-brain volume as well as significant dropout in the orbitofrontal cortex for some participants, group-level analyses used a mask that did not include the most dorsal part of the parietal lobe, and most areas in the orbitofrontal cortex (BA11) and the ventral frontal pole.

Preprocessing and analysis of imaging data were performed using Statistical Parametric Mapping software (SPM8; Wellcome Department of Imaging Neuroscience, Institute of Neurology, London UK) as well as custom code written in Matlab (MathWorks). Preprocessing of EPI images included high-pass filtering of the data with a 128 Hz filter, motion correction (rigid-body realignment of functional volumes to the first volume), coregistration to MNI atlas space to facilitate group analysis (by computing an affine transformation of the structural images to the functional images, and then to the MNI template, segmentation of the structural image for nonlinear spatial normalization, and finally nonlinear warping (i.e., normalization) of both functional and structural images), and spatial smoothing using a Gaussian kernel with a full-width at half-maximum of 8 mm, to allow for statistical parametric mapping analysis. Statistical analyses of functional time series followed both a model-based and a model-agnostic approach. Structural images were averaged together to permit anatomical localization of functional activations at the group level.

**Model-based analysis.** In the model-based analysis, we used the best-fitting computational model to generate a set of neural hypotheses that took the form of predictions for the specific time courses of internal variables of interest. For this, we fit the model to each participant's data from the functional scans (300 trials per participant). We then used the maximum a posteriori parameters to run the model and generate variables of interest: the weights of each of the features and the prediction error at the time of the outcome for each trial. We analyzed the whole-brain BOLD data using a general linear model (GLM) that included the following two regressors of interest: (1) the prediction error, that is, the difference between the obtained outcome and the (model-generated) value of the chosen stimulus on each trial; and (2) the standard deviation of the weights of the features of the chosen stimulus on each trial. The prediction error parametric regressor modulated outcome onsets, while the standard deviation parametric regressor modulated stimulus onsets. In addition, the GLM included regressors of no interest, as follows: (1) a stick regressor for the onsets of all stimuli; (2) a stick regressor and the onsets of all outcomes; (3) a block regressor spanning the duration between stimulus onset and the time the response was registered, to control for activity that correlates with longer reaction times; (4) a parametric regressor at the time of stimulus onset corresponding to the reaction time on that trial, to additionally account for activity that can be explained by the difficulty or amount of deliberation on each trial (Grinband et al.,

2008); and (5) six covariate motion regressors. None of the parametric regressors were orthogonalized to each other so that the variance that is shared between two regressors would not be attributed to either of them. Separate regressors were defined for each of the four runs. Each participant's data were then regressed against the full GLM, and coefficient estimates from each participant were used to compute random-effects group statistics at the second level. One contrast was tested for each of the regressors of interest to identify activity correlated with that regressor.

**Neural model comparison.** In a second, model-agnostic whole-brain analysis, a GLM was created that included all of the regressors of no interest cited above, and one regressor of interest: a parametric regressor at the time of stimulus onset that increased linearly from 0 to 1 across the trials of the game. This regressor was used to identify areas that are more active in the beginning of the game compared with the end of the game (i.e., areas that are inversely correlated to the regressor) as candidates for areas that are involved in representation learning. We used a voxel-level threshold of  $t = 4.78$  ( $p < 5 \times 10^{-5}$ ) combined with FWE cluster-level correction of  $p < 0.01$  to extract functional regions of interest (ROIs) for neural model comparison: a region in the right intraparietal cortex comprising 128 voxels [peak voxel (MNI coordinates), [33, -79, 40];  $t_{(21)} = 7.30$ ]; a region in the left intraparietal cortex comprising 123 voxels (peak voxel, [-24, -64, 43];  $t_{(21)} = 6.15$ ); a region in the precuneus comprising 95 voxels (peak voxel, [-3, -73, 46];  $t_{(21)} = 6.70$ ); a region in the right middle frontal gyrus (BA9) comprising 67 voxels (peak voxel, [39, 23, 28];  $t_{(21)} = 4.71$ ); a region in the left middle frontal gyrus (BA9) comprising 45 voxels (peak voxel, [-39, 2, 34];  $t_{(21)} = 4.58$ ); and a large region of activity (1191 voxels, spanning both hemispheres) in the occipital lobe, including BA17–B19, the fusiform gyrus, cuneus, and lingual gyrus, and extending bilaterally to the posterior lobe of the cerebellum (peak voxel, [3, -94, 7];  $t_{(21)} = 10.29$ ).

For each ROI and each participant, we extracted the time courses of BOLD activity from all voxels in the ROI and used singular value decomposition to compute a single weighted average time course per participant per ROI. We then removed from these time courses all effects of no interest by estimating and subtracting from the data, for each session separately, a linear regression model that included two onset regressors for stimulus and outcome onsets, a parametric regressor at stimulus onset corresponding to the reaction time on that trial (for valid trials only) and a block regressor on each valid trial that contained 1s throughout the duration of the reaction time, six motion regressors (3D translation and rotation), two trend regressors (linear and quadratic), and a baseline. All regressors, apart from the motion, trend, and baseline regressors, were convolved with a standard hemodynamic response function (HRF) before being regressed against the time-course data.

We used the residual time courses to compare the five models that made predictions for attention on each trial (i.e., all models except the naïve RL model). For each model, we created a regressor for the degree of representation learning/attentional control at each trial onset, as follows: the standard deviation of weights of the chosen stimulus for the fRL and fRL + decay models; the standard deviation of  $\phi$  for the hybrid model; the standard deviation of the inferred probability that the participant is testing each of the hypotheses on this trial for the SH model; and the standard deviation of the probability that each of the features of the chosen stimulus is the target feature for the Bayesian model. We then computed the log likelihood of a linear model for the neural time course containing this regressor convolved with a standard HRF. Since linear regression provides the maximum likelihood solution to a linear model with Gaussian-distributed noise, the maximum log likelihood of the model can be assessed as follows:

$$\mathcal{LL} = -N_{\text{data}} \cdot \left[ \ln \left( \sqrt{2\pi\hat{\sigma}^2} \right) + 0.5 \right], \quad (15)$$

where  $N_{\text{data}}$  is the total number of data points in the time-course vector, and  $\hat{\sigma}$  is the standard deviation of the residuals after subtracting the best-fit linear model. Since all models had one parameter (the coefficient of the single regressor), their likelihoods could be directly compared to ask which model accounted best for the neural activation patterns. All

neural model comparison code was developed in-house in Matlab and is available on-line at [www.princeton.edu/~nivlab/code](http://www.princeton.edu/~nivlab/code).

## Results

Participants played short games of the dimensions task—a probabilistic multidimensional bandit task where only one dimension is relevant for determining reward at any point in time—with the relevant dimension (and within it, the target feature, which led to 75% reward) changing after every game. Figure 1*B* depicts a sequence of choices in one game. In this example, the participant learned within 10 trials that the reward-relevant dimension is “color” and the target feature is “yellow.” One might also infer that the participant initially thought that the target feature might be “circle” on the “shape” dimension. It is less clear whether the participant later entertained the hypothesis that plaid texture is the target feature (and, in fact, whether the wavy texture was chosen on the first four trials purposefully or by chance). The overall learning curve across participants and games is shown in Figure 1*C*. On average, participants learned the task; however, their performance at the end of games was far from perfect (~60% correct). Examination of the number of correct choices on the last six trials of each game revealed that, indeed, participants learned only ~40% of the games (Fig. 1*D*). Note that the occurrence of “unlearned games” is beneficial, as in these games we can analyze the learning process throughout the whole game (performance after learning has been completed is less interesting for our intentions).

The game segment depicted in Figure 1*B* illustrates both the richness of the data from our task, in which a sequence of 10 choices may be interpreted as involving testing of (at least) three hypotheses, as well as its paucity—on any given trial, we know what stimulus the participant chose, but not why they chose it. That is, we do not know what feature or combination of features led the participant to choose this stimulus. This difficulty in interpreting choices in our task, together with our interest in understanding the nature of the learning dynamics, motivated a model-based analysis of the behavioral data.

## Modeling the dynamics of learning

We fit to the behavioral choice data (available online at [www.princeton.edu/~nivlab/data](http://www.princeton.edu/~nivlab/data)) a series of computational models ranging from suboptimal to statistically optimal solutions for the task (Fig. 2*A*). These models embody different hypotheses regarding the effect of the outcome of one trial on subsequent representations and, consequently, on choices. In essence, each hypothesis (model) consists of a sequence of predictions for the participant's choice on the next trial, given the trials since the beginning of the game. Specifying these hypotheses as computational models allowed us to test them rigorously and compare them to each other in a statistically principled way.

We first tested two benchmark models: a naïve RL model that does not take advantage of the structure of the task and of the knowledge that only one dimension is relevant for determining reward (this model serves as a baseline for learning performance absent any representation learning); and a Bayesian model that solves the task in a statistically optimal way. The naïve model learned a separate value for each of the 27 possible stimuli. That is, in the naïve model, receiving a point for choosing the green square with plaid texture led to an increased probability of selecting green squares with plaid texture. However, the model could not generalize from this experience to other green stimuli or to other squares. We did not expect this model to perform well on the task or to provide a good explanation for participants' behav-

ior—with only 20 trials per game on average, there was not even sufficient time to sample each of the 27 stimuli. Nevertheless, this model was better than chance at predicting participants' choices (mean likelihood per trial for left-out data, 0.401;  $t_{(21)} = 10.65$ ,  $p < 10^{-9}$ , Student's  $t$  test comparison with the random baseline; Fig. 2A, top).

In contrast to the naïve RL model, the Bayesian model exploits all of the available knowledge about the structure of the task. Nevertheless, this model also did rather poorly in explaining the data, with an average likelihood per trial of 0.408 (Fig. 2A, bottom). In fact, the Bayesian model was not significantly better than the naïve RL model at predicting participants' choices ( $t_{(21)} = 0.75$ ,  $p = 0.46$ ; see also Fig. 2B, bottom two curves), despite embodying the statistically optimal solution of the task. This is perhaps not surprising, even in light of the recent burgeoning literature on Bayesian inference in perception, given that a full Bayesian solution to representation learning is computationally not feasible in realistic multidimensional environments—Bayesian inference was tractable in our task only due to the small number of stimuli and dimensions, and the fact that the reward probabilities were known to be stable a priori. Hence, it is reasonable that, in general, the brain does not solve multidimensional RL tasks using Bayesian inference.

We then tested a series of models that embody different approximate solutions to reducing the dimensionality of the problem being learned. First, we tested the fRL model—a straightforward extension to the naïve model that takes advantage of the fact that, in our task, different features of a stimulus do not interact to determine reward. This model learns values (weights) for each feature rather than for combinations of features, such that obtaining 1 point for selecting the green square with plaid texture would cause the model to predict a higher likelihood of choosing green stimuli, square stimuli, or plaid stimuli in the future, with these effects being additive (i.e., the model predicts an even higher likelihood of choosing a green square if such a stimulus is available in the subsequent trial). The fRL model predicted participants' performance quite well (average likelihood per trial, 0.470; significantly better than either the Bayesian or the naïve RL models:  $t_{(21)} = 7.93$ ,  $p < 10^{-7}$  and  $t_{(21)} = 14.85$ ,  $p < 10^{-11}$ , respectively; Fig. 2A, second from top). This can be seen as reflecting the fact that, following our instructions and the structure of the task, participants learned at the level of features and not whole stimuli. It also suggests that, at the level of features, participants' learning process was more reminiscent of reinforcement learning than Bayesian learning (which also operated at the level of features). This notwithstanding, the fRL model treats all dimensions of the task as equal, and weighs the three features of a stimulus equally both in choice (i.e., in determining the value of a stimulus) and in learning (the prediction error is divided equally among the features), and thus the model does not fully take advantage of the known structure of the task.

In an attempt to improve upon the fRL model and allow it to differentially (and sensibly) weigh the contributions of features on different dimensions, we next tested a hybrid Bayesian-RL model. This model learns relevance weights for each dimension using Bayesian inference, and uses these to weight learning and choice in an fRL model. We parameterized this model such that Bayesian dimension weights could be used as is, accentuated (in the extreme, only the maximum a posteriori dimension would be taken into account) or relaxed toward equal weights. In this formulation, the fRL model is nested within the hybrid model (setting  $\alpha = 0$  in the hybrid model recovers the fRL model; see

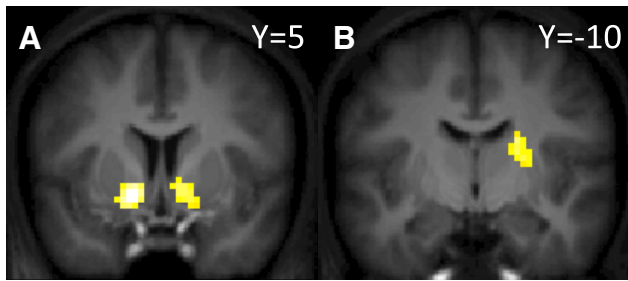
Materials and Methods). Interestingly, despite the addition of optimal inference at the level of dimension relevance, this model did not do better than the simple fRL model in predicting participants' performance (average likelihood per trial, 0.471;  $t_{(21)} = 0.39$ ,  $p = 0.7$ , when compared to the fRL model; Fig. 2A). Indeed, the best-fitting value of the parameter  $\alpha$  was, on average, close to 0 (see Table 1), the value at which the two models are identical (note that the curve for the hybrid model in Figure 2B lies behind that for the fRL model).

Another method of feature-level learning is the SH model developed by Wilson and Niv (2011). According to this model, participants test a single feature at each point in time (see Materials and Methods), serially searching for the target feature. This model is similar to the hybrid model in that it weighs dimensions differentially during both learning and choice. It is more extreme than the hybrid model as it concentrates on only one dimension (and, in fact, on one feature) at each point in time, but it is also less optimal than the hybrid model in that the feature being tested is chosen randomly (see Materials and Methods). Nevertheless, this model predicted participants' data slightly better than the fRL and hybrid models (average likelihood per trial, 0.493; significantly better than the fRL and hybrid models,  $t_{(21)} = 3.86$ ,  $p < 0.001$  and  $t_{(21)} = 3.47$ ,  $p < 0.005$ , respectively; Fig. 2A, second from bottom). Notably, this model does not use reinforcement learning at all.

The similar performances of the fRL, hybrid, and SH models, despite their rather radically different algorithms (equal weighting of dimensions vs learning about only a single feature; reinforcement learning vs likelihood ratio tests) suggested that the different models may be capturing nonoverlapping aspects of participants' choice and learning processes. Moreover, the poor performance of the Bayesian model, and the fact that uniform weighting of dimensions was preferable to Bayesian dimension weights in the hybrid model (Table 1), indicated that perhaps the models were doing "too well" at the task—they were outperforming the participants, and thus failing to predict participants' choices.

One difference between our models and human participants is that the models do not forget. Each of the models described above, whether based on statistical inference or RL, used all trials from the beginning of a game effectively. In particular, what the model learned about, say, the triangle feature, remained faithfully in memory even if the next few trials involved choices of stimuli with a circle shape. But this might not be the case in the human brain. To test whether forgetting of the learned values of features of unchosen stimuli could improve the correspondence between the choices of the model and human behavior, we developed the fRL+decay model, which learns weights for features using RL with uniform dimension weights (as in the fRL model) but decays ("forgets") the weights of unchosen features toward 0 [a similar Bayesian plus decay model was developed (Wilson and Niv, 2011), in which the posterior distribution over which feature is the target feature decays to a uniform baseline on every trial; this model performed significantly but only slightly better than the regular Bayesian model, and was still inferior to the other approximate models].

As can be seen in Figure 2A, the fRL+decay model provided the best fit for the data (average likelihood per trial, 0.528) and was significantly better than the fRL model ( $t_{(21)} = 15.33$ ,  $p < 10^{-12}$ , paired  $t$  test) as well as all other models. These results were similar when predicting the fast-paced trials or the slow-paced trials. Furthermore, fitting initial values for all reinforcement learning models and the target for decay for the



**Figure 3.** Neural correlates of prediction errors from the fRL+decay model. Activations were thresholded at a whole-brain FWE threshold of  $p < 0.05$  (which corresponded to  $t > 6.4$  and  $p < 1.5 \times 10^{-6}$  at the single-voxel level) and a minimum cluster size of 10 voxels. **A**, Activations in bilateral ventral striatum (left: peak MNI coordinates,  $[-15, 5, -11]$ ; peak intensity,  $t = 10.01$ ; cluster size, 57 voxels; right: peak MNI coordinates  $[12, 8, -8]$ ; peak intensity,  $t = 8.37$ ; cluster size, 47 voxels). **B**, Activation in dorsal putamen (peak MNI coordinates,  $[21, -7, 10]$ ; peak intensity,  $t = 8.55$ ; cluster size, 45 voxels). No other areas survived this threshold. Overlay: average structural scan of the 22 participants.

fRL+decay model did not significantly improve the fit of the fRL+decay model and did not change its superiority compared with other models.

Predicting participants' choices with  $>50\%$  accuracy in a task with three options (chance, 33%) may seem a modest success; however, at the beginning of games participants' choices cannot be predicted as all features have similar initial weights, and the longer the game, the better the model can predict choice. Figure 2*B* indeed shows that the predictive accuracy of the fRL+decay model increased, on average, to  $>80\%$  throughout the game, with both the fRL model and the Bayesian model significantly lagging behind (fRL+decay predictions were significantly better than those of the fRL model on all trials except the first and the 24th, and were significantly better than those of the Bayesian model on all trials except for the first trial in a game;  $p < 0.05$ , paired Student's  $t$  tests for each trial separately).

Moreover, as a more stringent test of the models' ability to predict participants' choices, we repeated this analysis considering only games in which the participant did not learn the identity of the target feature by the end of the game (i.e., games in which the participant chose the stimulus containing the target feature fewer than five times in the last six trials of the game). In these games, although participants were performing at chance (data not shown), the fRL+decay model could still predict choices with  $>70\%$  accuracy, on average, by the end of the game (Fig. 2*C*), with the other models significantly lagging behind. In particular, the Bayesian model could not predict participants' choices significantly better than chance in six of the last seven trials ( $p > 0.05$ ), further suggesting that this model could account for participants' behavior only to the extent that that behavior accorded with the correct solution of the task.

### Neural substrates for representation learning

We next looked for neural correlates of the learning process that is realized by the fRL+decay model. To start from well trodden ground, we first tested for correlates of prediction errors, that is, for the difference between the outcome (0 or 1 point) and the value of the chosen stimulus (as learned by the model) on every trial. As expected (McClure et al., 2003; O'Doherty et al., 2003, 2004; Hare et al., 2008; Diuk et al., 2010; Niv et al., 2012), a regressor for prediction errors correlated with BOLD signals in the striatum, specifically bilateral ventral striatum (nucleus accumbens) and right dorsal putamen (Fig. 3).

Having confirmed that the values that our model learns generate prediction errors that are in line with the known locations of

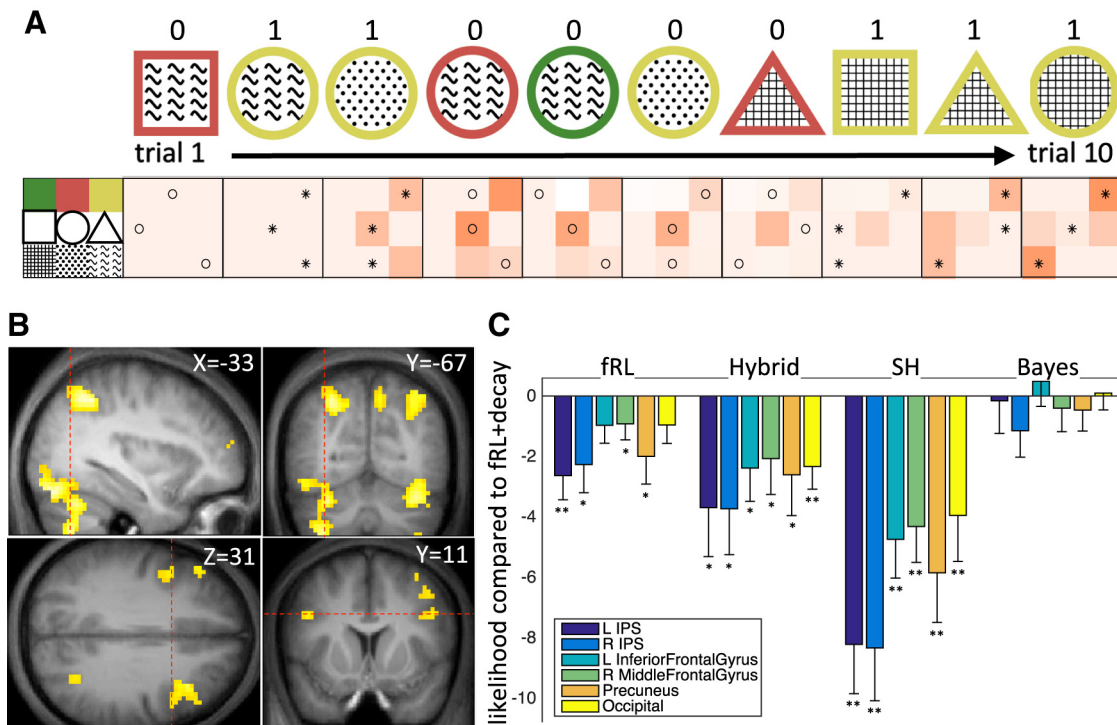
BOLD prediction-error activity, we next sought to investigate the correlates of representation learning, that is, the process by which participants homed in on the reward-relevant dimension. Contrary to our a priori intuition, which suggested that participants search for the relevant dimension and the target feature through a process of elimination (as embodied in, for instance, the SH model), model comparison had suggested that participants were simultaneously learning about all dimensions of the chosen stimuli. Indeed, the fRL+decay model does not require any specialized representation learning mechanism, and could conceivably be implemented solely within the RL mechanisms of the basal ganglia: dopaminergic prediction errors there could modulate the strengthening or weakening of corticostriatal synapses representing the features of the currently chosen stimulus (Wickens and Kötter, 1995; Reynolds et al., 2001; Wickens et al., 2003), with synapses corresponding to features of unchosen stimuli decaying uniformly as in a passive process of forgetting. Nevertheless, we asked whether we could find correlates of the representation learning process outside the basal ganglia.

To search for correlates of representation learning per se, we sought to quantify for each trial the degree to which the participant is engaged in representation learning. For instance, at the beginning of a game an area involved in representation learning would be required, whereas at other times in the game representation learning may have already terminated and the participant is simply selecting what she believes to be the target feature. Such a measure can be derived from the distribution of the nine feature weights (Fig. 4*A*): when these are fairly uniform (Fig. 4*A*, trials 1, 2, and 8), the participant must actively engage in representation learning, whereas when the participant homes in on certain features and selects them repeatedly, the weights of these features increase, thereby differentiating from the other weights that decay to 0. Importantly, this differentiation of weights occurs whenever features are consistently chosen (as in a choice kernel; Lau and Glimcher, 2005), and regardless of whether the consistently chosen feature is the target feature or not (Fig. 4*A*, compare the weights on trial 4, where the participant is choosing circles, to those on trial 10, where the participant has homed in on plaid or yellow—in both cases, two features have higher values than all others). We thus used the differentiation of feature weights (e.g., quantified as the standard deviation of the weights of the features of the chosen stimulus; see Materials and Methods) to search for a neural representation learning mechanism that is most active when all weights are similar, decreasing activity as participants focus on a certain feature and cease to consider alternatives.

We found significant correlates in areas approximately corresponding to what has been labeled “the frontoparietal attention control network” (Culham and Kanwisher, 2001; Corbetta and Shulman, 2002): bilateral intraparietal sulcus (IPS) and precuneus (Fig. 4*B*, top), and bilateral dorsolateral prefrontal cortex (dlPFC; specifically, activations in middle and inferior frontal gyrus on the left side and extending into the lateral orbitofrontal cortex in the right side; Fig. 4*B*, bottom). Additionally, we found strong activations in the occipital lobe including Brodmann areas 17, 18, and 19, the fusiform area, and extending to the posterior lobes of the cerebellum (Table 2), areas that have also been identified as part of the covert attention network (Corbetta et al., 1998; Corbetta and Shulman, 2002).

These results suggest that the frontoparietal attention control network is active in dynamically directing attention to different visual dimensions of the task stimuli during the representation learning phase, that is, when different alternatives for the relevant dimension are being entertained. In contrast, when the partici-





**Figure 4.** Neural substrates of representation learning. **A**, Sequence of choices and associated feature weights from the fRL + decay model. Weights for each of the nine task features (left) are depicted in the matrix under the chosen stimulus, with darker orange corresponding to a larger weight. Dots (filled for rewarded choices, empty for choices that led to 0 points) denote the three features chosen on the current trial; weights reflect estimates based on previous trials, before learning from the current trial, that is, the weights are the basis for the current choice, as per the model. **B**, Brain areas inversely correlated with the standard deviation of the weights of the chosen stimulus, at the time of stimulus onset. These areas are more active when weights are more uniform, as in trials 1, 2, and 8 above. Positive activations, thresholded at a  $p < 0.0001$  ( $t > 4.49$ ) voxelwise threshold and then subjected to a whole-brain FWE cluster-level threshold of  $p < 0.05$ , were significant in nine areas (Table 2). Shown here are bilateral IPS and precuneus (top), bilateral dlPFC (bottom), and bilateral occipital/cerebellar activations. Overlay: average structural scan of the 22 participants. Red dashed line, Slice coordinates. **C**, Neural model comparison. BOLD activity in six ROIs (identified using a model-agnostic GLM) supported the fRL + decay model when compared with the fRL, hybrid, and SH models, and was agnostic regarding the comparison between the fRL + decay model and the Bayesian model (the naïve RL model was not tested as it did not predict attentional control). Bars denote the log likelihood of each model minus that of the fRL + decay model, averaged across participants. Negative values represent higher log likelihood for the fRL + decay model. Error bars denote SEM. **\*\*** $p < 0.01$ , **\*** $p < 0.05$ , one-tailed paired Student's  $t$  test.

**Table 2.** Brain areas that were significantly correlated with an “attention” regressor (BOLD activity anticorrelated with the standard deviation of feature weights of the chosen stimulus), which passed a whole-brain cluster correction threshold of  $p < 0.05$  (voxel threshold:  $p < 0.001$ ,  $t > 4.49$ )

Brain area	Cluster size (voxels)	Peak MNI coordinates (x, y, z)	Peak intensity ( $t_{(21)}$ value)
Left intraparietal cortex	300	(−27, −64, 43)	7.77
Right cuneus/inferior occipital gyrus/fusiform and extending to cerebellum	399	(9, −94, 4)	7.45
Left fusiform and extending to cerebellum	534	(−36, −64, −50)	7.03
Right dorsolateral prefrontal cortex extending into lateral orbital cortex	270	(36, 26, 28)	6.75
Right temporal lobe	44	(57, −46, −14)	6.20
Right intraparietal cortex	200	(33, −61, 46)	5.87
Right precuneus	35	(9, −70, 52)	5.53
Left middle frontal gyrus	20	(−48, 32, 31)	5.15
Left inferior frontal gyrus	23	(−42, 5, 34)	5.14

pant concentrates on one feature (Zhou and Desimone, 2011; Baldauf and Desimone, 2014), this system is less active. Importantly, our analysis accounted for variance related to stimulus and outcome onsets, prediction error activity at the time of outcome, and reaction times in regressors of no interest (see Materials and Methods), suggesting that these activations cannot simply be attributed to task difficulty, reward rate, or surprise.

Finally, we used neural model comparison to ask whether the neural activations in the attentional control network can arbitrate between competing models of the underlying representation learning process (Mack et al., 2013). To do this, we first extracted activations in the attentional-control network using a model-agnostic method by searching for areas that were more active in the beginning of games (when representation learning was heavily called for) and had reduced activity as the game wore on. This contrast identified six ROIs corresponding to the frontoparietal attention control network (a single occipital region spanning both hemispheres and extending to cerebellum, bilateral IPS, precuneus, and bilateral dlPFC; see Materials and Methods). We then extracted and averaged the signals in each of these areas to create six time courses per participant, and modeled and subtracted from these data all factors of no interest. The residual signal was then modeled separately for each of the competing models, as a linear function of the “attentional control” predicted by the model (see Materials and Methods).

Results of the neural model comparison favored the fRL + decay model (Fig. 4C) more than each of the other models, apart from the Bayesian model, which provided a similarly good fit to the neural data. That is, the likelihood of a model that generates the BOLD signals in left IPS, right IPS, or precuneus ROIs from the attentional control measure derived from the fRL + decay process was significantly greater than the likelihood of a model generating these BOLD signals from attentional control as derived from the fRL model (Fig. 4C). Similarly, model

evidence for fRL+decay was significantly greater than that of the hybrid and SH models in all six ROIs. Surprisingly, model evidence for the Bayesian model was not significantly different from that of the fRL+decay model in any of the ROIs (Fig. 4C). However, we note that the Bayesian plus decay model mentioned briefly above, which provided a superior fit to the behavioral data compared with the pure Bayesian model, showed a significantly worse fit to the neural data as compared with the fRL+decay model in both the right IPS and the precuneus. Moreover, when testing the fit of the model predictions to neural data from unlearned games only, the fRL+decay model fit all ROIs significantly better than the Bayesian model (data not shown). This suggests that the good fit of the Bayesian model to the neural data, as seen in Figure 4C, might reflect a good fit to the learned phase of games, where attention is more focused than the fRL+decay model might allow, and is more in line with the Bayesian model. In contrast, during the learning phase only, the fRL+decay model accounts for the neural data better than does the Bayesian model. These neural results, together with the behavioral support for the fRL+decay model, suggest that the fRL+decay model captures important aspects of the representation learning process and its neural control.

## Discussion

To study the interaction between representation learning and reinforcement learning, we tested participants on the dimensions task—a multidimensional three-armed bandit task with probabilistic rewards. By comparing alternative computational models for the learning process, we showed that, rather than using a Bayes-optimal strategy to solve the task, participants combined value learning at the level of features with a decay process that served to highlight the value of consistently chosen features and decay the value of other features to 0. This fRL+decay model explained up to 70% of the choice variance even in unlearned games—those games in which participants performed randomly with respect to choosing the target feature—suggesting that the model captured meaningful aspects of the underlying learning and decision-making process. Neural analysis using a model-derived trial-by-trial measure of the extent to which participants were engaged in representation learning implicated the frontoparietal attentional control network in representation learning. This bilateral network, which includes the IPS, precuneus, and dlPFC, was more active when participants were engaged in representation learning and had not yet homed in on one feature that they believed was the target feature. Extrapolating to tasks in the real world, our results suggest that trial-and-error learning in multidimensional environments is aided by neural attentional control mechanisms that help to focus learning on a subset of dimensions, thus mitigating the curse of dimensionality and enabling efficient learning.

Although no two experiences are exactly alike, to make adaptive decisions we must learn from past experience. As such, learning is predicated on generalization—deciding what details are inconsequential and can be ignored so as to allow prediction in one situation based on experience in another (Shepard, 1987; Jones and Cañas, 2010). In RL, function approximation methods allow for generalization by assuming smoothness over the values of similar states (e.g., the learned value for a dark red stimulus must be similar to that of a light red stimulus, a sensible approximation given that the difference in color may arise from observing the same stimulus in different lighting conditions). However, a problem still remains when attempting to generalize across dimensions, for instance, from red to triangle, where similarity

functions are not well defined. In fact, by definition generalization occurs within, but not across “separable” (as opposed to “integral”) perceptual dimensions such as color, shape, and texture (Nosofsky and Palmeri, 1996; Soto et al., 2014). Thus, to overcome the curse of dimensionality one must reduce the number of relevant dimensions, effectively learning from experience a minimal state representation for each task (Jones and Cañas, 2010). Here we have investigated this learning process computationally and neurally in a task that involves reducing the dimensionality of the task from three dimensions to one.

Obviously, our task was highly simplified. Nevertheless, we believe that only a few dimensions are relevant to most naturalistic tasks. This is essentially an assumption about the complexity of the causal structure of tasks that we are faced with in everyday life (or at least tasks that we are good at solving). Our hypothesis is that brain mechanisms have evolved to take advantage of such regularities across tasks, in essence developing a prior on task structure that makes selective attention beneficial. This prior was built into the dimensions task explicitly through task instructions; however, work in multidimensional categorization tasks suggests that, even in the absence of instructions, animals and humans assume a simple structure in which category structure depends on only one dimension, and test more complex hypotheses only once they have exhausted all such simple hypotheses (Shepard et al., 1961; Smith et al., 2004; Rehder and Hoffman, 2005b; Cohen and Schneiderman, 2013).

Indeed, considerable behavioral work on categorization and concept formation has explored how humans learn what dimensions of a stimulus are relevant (for review, see Ashby and Maddox, 2005), with selective attention implicated as a key factor in multidimensional learning (Nosofsky, 1986; Kruschke, 1992, 2006; Nosofsky et al., 1994a; Rehder and Hoffman, 2005a). This work has also demonstrated that the selective nature of attention is adaptive, with selectivity developing only in tasks in which the category structure depends on only a few of the stimulus dimensions. Selective attention is thus an integral part of most categorization theories: learned attention weights amplify or attenuate specific stimulus dimensions to facilitate category discrimination.

Two general classes of models have been suggested for category learning. On the one hand are rule-learning models that use sequential hypothesis testing (Nosofsky et al., 1994b). These models suggest that hypotheses regarding the rule for categorization are tested from simple to complex, with attention focusing first on one dimension at a time (Levine, 1959, 1969; Bower and Trabasso, 1963; Trabasso and Bower, 1964), and widening only as needed. On the other hand are selective attention models that use either a fixed allocation of attention across dimensions (Nosofsky, 1986) or adjust this allocation dynamically (Kruschke, 1992). The latter models suggest that attention is broad at first, and the focus of attention gradually narrows. Empirical evidence supports both processes (Rehder and Hoffman, 2005a), as has been captured by more recent models (Love et al., 2004).

The models we have tested here can be construed as adaptations of models from the categorization literature to a probabilistic RL scenario. In particular, our data support the fRL+decay model, which is, in a sense, a hybrid between serial hypothesis testing and a parallel learning model that narrows the focus of attention only gradually. Although this model does not explicitly involve attention processes or hypothesis testing, it emulates both: the decay of weights of unchosen features allows the model to focus learning on the weight of one consistently chosen feature. At the same time, the decay implements a form of a “choice kernel” that allows the model to better predict future choices

based on the repetition of actions at the level of features (Lau and Glimcher, 2005; Schönberg et al., 2007; Wilson and Niv, 2011; Seymour et al., 2012; Akaishi et al., 2014). However, the superior performance of the fRL+decay model cannot be wholly attributed to a choice kernel, as simply adding a feature-level choice kernel to the fRL model (without weight decay) improved the fit of the model compared with the fRL model ( $p < 10^{-7}$ , paired  $t$  test), but was still inferior to the fRL+decay model ( $p < 10^{-9}$ , paired  $t$  test; results not shown). Thus, we can conclude that although the fRL+decay maintains an implicit choice kernel, that is not the sole reason that it accounts for the behavioral data as well as it does.

Because the fRL+decay model dynamically adjusts the “width” of its choice kernel according to previous choices, it can also be seen as adjusting the breadth of attention: if red triangles are available on several trials, and are consistently chosen, the model will learn equally about both features. That having been said, the fRL+decay model learns equally about all features of a chosen stimulus, and later “unlearns” the values of those features that are not consistently chosen. One could imagine that selective attention would act at the time of learning, thus obviating the need for later forgetting. It would be interesting in future work to devise a version of the task in which selective attention could be measured explicitly (e.g., using eye gaze; Rehder and Hoffman, 2005a) and to test whether attentional weighting at learning can replace the decay process.

Previous work in reinforcement learning addressing related questions of learning a representation (Doya et al., 2002; Frank and Badre, 2012) used a “mixture-of-experts” architecture to infer what stimulus dimensions participants were attending to. In that framework, each “expert” learns a behavioral policy based on a subset of stimulus features, and the action recommendations of all experts are combined using “attention weights” that correspond to how much reward each expert has led to in the past. An implementation of this framework for our task, with three experts each learning a choice policy for one dimension, is conceptually straightforward. In effect, our hybrid model can be seen as an instance of such an architecture, with weights in each dimension determining the policy for one expert, and the Bayesian dimension weights setting the responsibilities of each expert. This specific model did not provide a good explanation of our data, but other mixture-of-experts variants might perform better. In any case, it is unlikely that the brain uses a full mixture-of-experts architecture for representation learning as this would necessitate experts for each combination of environmental dimensions (Frank and Badre, 2012), bringing back the curse of dimensionality in complex environments.

In contrast to the algorithmic basis of representation learning, the neural basis of this process has been explored less often—what brain mechanisms are involved in deciding what aspects of the environment should RL structures such as the striatum learn values for? Our results point to both the IPS, an area in the dorsal posterior parietal cortex implicated in endogenous attention processes and in visual feature search among multidimensional stimuli (Culham and Kanwisher, 2001; Chica et al., 2011; Liu et al., 2011; Wei et al., 2011) and the dlPFC, a frontal area associated with attentional control, in particular in the interaction with posterior parietal cortex, and in switching between different tasks (Dove et al., 2000; Corbetta and Shulman, 2002). Much work suggests that the dlPFC mediates shifts in attention between perceptual features of complex stimuli according to task demands, as is required, for instance, in the WCST and its animal analogs (Owen et al., 1991; Dias et al., 1996a,b, 1997; Birrell and Brown,

2000; Dalley et al., 2004; Buchsbaum et al., 2005; Fletcher et al., 2005; Floresco et al., 2008). Prefrontal dopamine has also been implicated in the selective attention processes thought to underlie such “set formation” and “set shifting” (Miller and Wickens, 1991; Holland and Gallagher, 1999; Crofts et al., 2001; Chudasama and Robbins, 2004). Our study was not specifically designed to assess the contribution of reward prediction errors, putatively conveyed by dopaminergic projections, to the function of the dlPFC and IPS—this is another promising venue for future work.

In sum, while reinforcement learning methods are notorious for not scaling up to tasks of real-world complexity, our results suggest that selective attention mechanisms that direct attention to only a subset of environmental dimensions at each point in time can help to mitigate the curse of dimensionality. This interaction between attention and learning can vastly increase the efficiency of simple trial-and-error learning processes in the basal ganglia, thus moving one step toward accounting for the amazing adaptive capabilities of humans, even in the face of a dynamic, multidimensional world.

## References

- Akaishi R, Umeda K, Nagase A, Sakai K (2014) Autonomous mechanism of internal choice estimate underlies decision inertia. *Neuron* 81:195–206. [CrossRef Medline](#)
- Ashby FG, Maddox WT (2005) Human category learning. *Annu Rev Psychol* 56:149–178. [CrossRef Medline](#)
- Baldauf D, Desimone R (2014) Neural mechanisms of object-based attention. *Science* 344:424–427. [CrossRef Medline](#)
- Bar-Gad I, Havazelet-Heimer G, Goldberg JA, Ruppel E, Bergman H (2000) Reinforcement-driven dimensionality reduction—a model for information processing in the basal ganglia. *J Basic Clin Physiol Pharmacol* 11:305–320. [Medline](#)
- Barto AG (1995) Adaptive critic and the basal ganglia. In: *Models of information processing in the basal ganglia* (Houk JC, Davis JL, Beiser DG, eds), pp 215–232. Cambridge, MA: MIT.
- Birrell JM, Brown VJ (2000) Medial frontal cortex mediates perceptual attentional set shifting in the rat. *J Neurosci* 20:4320–4324. [Medline](#)
- Bishara AJ, Kruschke JK, Stout JC, Bechara A, McCabe DP, Bussey JR (2010) Sequential learning models for the Wisconsin card sort task: assessing processes in substance dependent individuals. *J Math Psychol* 54:5–13. [CrossRef Medline](#)
- Bower G, Trabasso T (1963) Reversals prior to solution in concept identification. *J Exp Psychol* 66:409–418. [CrossRef Medline](#)
- Buchsbaum BR, Greer S, Chang WL, Berman KF (2005) Meta-analysis of neuroimaging studies of the Wisconsin card-sorting task and component processes. *Hum Brain Mapp* 25:35–45. [CrossRef Medline](#)
- Cañas F, Jones M (2010) Attention and reinforcement learning: constructing representations from indirect feedback. Paper presented at CogSci 2010: The Annual Meeting of the Cognitive Science Society, Portland, OR, August.
- Chica AB, Bartolomeo P, Valero-Cabré A (2011) Dorsal and ventral parietal contributions to spatial orienting in the human brain. *J Neurosci* 31:8143–8149. [CrossRef Medline](#)
- Chudasama Y, Robbins TW (2004) Dopaminergic modulation of visual attention and working memory in the rodent prefrontal cortex. *Neuropsychopharmacology* 29:1628–1636. [CrossRef Medline](#)
- Cohen Y, Schneidman E (2013) High-order feature-based mixture models of classification learning predict individual learning curves and enable personalized teaching. *Proc Natl Acad Sci U S A* 110:684–689. [CrossRef Medline](#)
- Corbetta M, Shulman GL (2002) Control of goal-directed and stimulus-driven attention in the brain. *Nat Rev Neurosci* 3:201–215. [CrossRef Medline](#)
- Corbetta M, Akbudak E, Conturo TE, Snyder AZ, Ollinger JM, Drury HA, Linenweber MR, Petersen SE, Raichle ME, Van Essen DC, Shulman GL (1998) A common network of functional areas for attention and eye movements. *Neuron* 21:761–773. [CrossRef Medline](#)
- Crofts HS, Dalley JW, Collins P, Van Denderen JC, Everitt BJ, Robbins TW,

- Roberts AC (2001) Differential effects of 6-OHDA lesions of the frontal cortex and caudate nucleus on the ability to acquire an attentional set. *Cereb Cortex* 11:1015–1026. [CrossRef Medline](#)
- Culham JC, Kanwisher NG (2001) Neuroimaging of cognitive functions in human parietal cortex. *Curr Opin Neurobiol* 11:157–163. [CrossRef Medline](#)
- Dalley JW, Cardinal RN, Robbins TW (2004) Prefrontal executive and cognitive functions in rodents: neural and neurochemical substrates. *Neurosci Biobehav Rev* 28:771–784. [CrossRef Medline](#)
- Daw ND (2011) Trial by trial data analysis using computational models. In: *Decision making, affect, and learning: attention and performance* (Delgado MR, Phelps EA, Robbins TW, eds), p xxiii. Oxford: Oxford UP.
- Dias R, Robbins TW, Roberts AC (1996a) Dissociation in prefrontal cortex of affective and attentional shifts. *Nature* 380:69–72. [CrossRef Medline](#)
- Dias R, Robbins TW, Roberts AC (1996b) Primate analogue of the Wisconsin Card Sorting Test: effects of excitotoxic lesions of the prefrontal cortex in the marmoset. *Behav Neurosci* 110:872–886. [CrossRef Medline](#)
- Dias R, Robbins TW, Roberts AC (1997) Dissociable forms of inhibitory control within prefrontal cortex with an analog of the Wisconsin Card Sort Test: restriction to novel situations and independence from “online” processing. *J Neurosci* 17:9285–9297. [Medline](#)
- Diuk C, Barto AG, Botvinick MB, Niv Y (2010) Hierarchical reinforcement learning: an fMRI study of learning in a two-level gambling task. *Soc Neurosci Abstr* 36:907.14.
- Dove A, Pollmann S, Schubert T, Wiggins CJ, von Cramon DY (2000) Prefrontal cortex activation in task switching: an event-related fMRI study. *Brain Res Cogn Brain Res* 9:103–109. [CrossRef Medline](#)
- Doya K, Samejima K, Katagiri K, Kawato M (2002) Multiple model-based reinforcement learning. *Neural Comput* 14:1347–1369. [CrossRef Medline](#)
- Fletcher PJ, Tenn CC, Rizos Z, Lovic V, Kapur S (2005) Sensitization to amphetamine, but not PCP, impairs attentional set shifting: reversal by a D1 receptor agonist injected into the medial prefrontal cortex. *Psychopharmacology (Berl)*, 183:190–200. [CrossRef Medline](#)
- Floresco SB, Block AE, Tse MT (2008) Inactivation of the medial prefrontal cortex of the rat impairs strategy set-shifting, but not reversal learning, using a novel, automated procedure. *Behav Brain Res* 190:85–96. [CrossRef Medline](#)
- Frank MJ, Badre D (2012) Mechanisms of hierarchical reinforcement learning in corticostriatal circuits 1: computational analysis. *Cereb Cortex* 22:509–526. [CrossRef Medline](#)
- Gershman SJ, Niv Y (2010) Learning latent structure: carving nature at its joints. *Curr Opin Neurobiol* 20:251–256. [CrossRef Medline](#)
- Gluck MA, Shohamy D, Myers C (2002) How do people solve the “weather prediction” task? Individual variability in strategies for probabilistic category learning. *Learn Mem* 9:408–418. [CrossRef Medline](#)
- Grinband J, Wager TD, Lindquist M, Ferrera VP, Hirsch J (2008) Detection of time-varying signals in event-related fMRI designs. *Neuroimage* 43:509–520. [CrossRef Medline](#)
- Hare TA, O’Doherty J, Camerer CF, Schultz W, Rangel A (2008) Dissociating the role of the orbitofrontal cortex and the striatum in the computation of goal values and prediction errors. *J Neurosci* 28:5623–5630. [CrossRef Medline](#)
- Holland PC, Gallagher M (1999) Amygdala circuitry in attentional and representational processes. *Trends Cogn Sci* 3:65–73. [CrossRef Medline](#)
- Jones M, Cañas F (2010) Integrating reinforcement learning with models of representation learning. Paper presented at CogSci 2010: The Annual Meeting of the Cognitive Science Society, Portland, OR, August.
- Kruschke J (2006) Learned attention. Paper presented at ICDL 2006: International Conference on Development and Learning—Dynamics of Development and Learning, Bloomington, IN, June.
- Kruschke JK (1992) Alcov: an exemplar-based connectionist model of category learning. *Psychol Rev* 99:22–44. [CrossRef Medline](#)
- Lau B, Glimcher PW (2005) Dynamic response-by-response models of matching behavior in rhesus monkeys. *J Exp Anal Behav* 84:555–579. [CrossRef Medline](#)
- Lawrence V, Houghton S, Douglas G, Durkin K, Whiting K, Tannock R (2004) Executive function and ADHD: a comparison of children’s performance during neuropsychological testing and real-world activities. *J Atten Disord* 7:137–149. [CrossRef Medline](#)
- Levine M (1959) A model of hypothesis behavior in discrimination learning set. *Psychol Rev* 66:353–366. [CrossRef Medline](#)
- Levine M (1969) Neo-noncontinuity theory. In: *The psychology of learning and motivation* (Bower GH, Spence JT, eds), pp 101–134. New York: Academic.
- Liu T, Hospadaruk L, Zhu DC, Gardner JL (2011) Feature-specific attentional priority signals in human cortex. *J Neurosci* 31:4484–4495. [CrossRef Medline](#)
- Love BC, Medin DL, Gureckis TM (2004) Sustain: a network model of category learning. *Psychol Rev* 111:309–332. [CrossRef Medline](#)
- Lyvers MF, Maltzman I (1991) Selective effects of alcohol on Wisconsin Card Sorting Test performance. *Br J Addict* 86:399–407. [CrossRef Medline](#)
- Mack ML, Preston AR, Love BC (2013) Decoding the brain’s algorithm for categorization from its neural implementation. *Curr Biol* 23:2023–2027. [CrossRef Medline](#)
- McClure SM, Berns GS, Montague PR (2003) Temporal prediction errors in a passive learning task activate human striatum. *Neuron* 38:339–346. [CrossRef Medline](#)
- Miller R, Wickens JR (1991) Corticostriatal cell assemblies in selective attention and in representation of predictable and controllable events. *Concepts Neurosci* 2:65–95.
- Milner B (1963) Effects of different brain lesions on card sorting. *Arch Neurol* 9:100–110.
- Monchi O, Petrides M, Doyon J, Postuma RB, Worsley K, Dagher A (2004) Neural bases of set-shifting deficits in Parkinson’s disease. *J Neurosci* 24:702–710. [CrossRef Medline](#)
- Montague PR, Dayan P, Sejnowski TJ (1996) A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *J Neurosci* 16:1936–1947. [Medline](#)
- Niv Y (2009) Reinforcement learning in the brain. *J Math Psychol* 53:139–154. [CrossRef](#)
- Niv Y, Schoenbaum G (2008) Dialogues on prediction errors. *Trends Cogn Sci* 12:265–272. [CrossRef Medline](#)
- Niv Y, Edlund JA, Dayan P, O’Doherty JP (2012) Neural prediction errors reveal a risk-sensitive reinforcement-learning process in the human brain. *J Neurosci* 32:551–562. [CrossRef Medline](#)
- Nosofsky RM (1986) Attention, similarity, and the identification-categorization relationship. *J Exp Psychol Gen* 115:39–61. [CrossRef Medline](#)
- Nosofsky RM, Palmeri TJ (1996) Learning to classify integral-dimension stimuli. *Psychon Bull Rev* 3:222–226. [CrossRef Medline](#)
- Nosofsky RM, Palmeri TJ, McKinley SC (1994a) Rule-plus-exception model of classification learning. *Psychol Rev* 101:53–79. [CrossRef Medline](#)
- Nosofsky RM, Gluck MA, Palmeri TJ, McKinley SC, Glauthier P (1994b) Comparing models of rule-based classification learning: a replication and extension of Shepard, Hovland, and Jenkins (1961). *Mem Cognit* 22:352–369. [CrossRef Medline](#)
- O’Doherty J, Dayan P, Schultz J, Deichmann R, Friston K, Dolan RJ (2004) Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science* 304:452–454. [CrossRef Medline](#)
- O’Doherty JP, Dayan P, Friston K, Critchley H, Dolan RJ (2003) Temporal difference learning model accounts for responses in human ventral striatum and orbitofrontal cortex during Pavlovian appetitive learning. *Neuron* 38:329–337. [CrossRef Medline](#)
- Ornstein TJ, Iddon JL, Baldacchino AM, Sahakian BJ, London M, Everitt BJ, Robbins TW (2000) Profiles of cognitive dysfunction in chronic amphetamine and heroin abusers. *Neuropsychopharmacology* 23:113–126. [CrossRef Medline](#)
- Owen AM, Roberts AC, Polkey CE, Sahakian BJ, Robbins TW (1991) Extra-dimensional versus intra-dimensional set shifting performance following frontal lobe excisions, temporal lobe excisions or amygdalo-hippocampectomy in man. *Neuropsychologia* 29:993–1006. [CrossRef Medline](#)
- Owen AM, Roberts AC, Hodges JR, Summers BA, Polkey CE, Robbins TW (1993) Contrasting mechanisms of impaired attentional set-shifting in patients with frontal lobe damage or Parkinson’s disease. *Brain* 116:1159–1175. [CrossRef Medline](#)
- Price AL (2009) Distinguishing the contributions of implicit and explicit processes to performance of the weather prediction task. *Mem Cognit* 37:210–222. [CrossRef Medline](#)
- Rehder B, Hoffman AB (2005a) Eyetracking and selective attention in category learning. *Cogn Psychol* 51:1–41. [CrossRef Medline](#)

- Rehder B, Hoffman AB (2005b) Thirty-something categorization results explained: selective attention, eyetracking, and models of category learning. *J Exp Psychol Learn Mem Cogn* 31:811–829. [CrossRef Medline](#)
- Rescorla RA, Wagner AR (1972) A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement. In *Classical conditioning II: current research and theory* (Black AH, Prokasy WF, eds), pp 64–99. New York: Appleton-Century-Crofts.
- Reynolds JN, Hyland BI, Wickens JR (2001) A cellular mechanism of reward-related learning. *Nature* 413:67–70. [CrossRef Medline](#)
- Rodriguez PF, Aron AR, Poldrack RA (2006) Ventral-striatal/nucleus-accumbens sensitivity to prediction errors during classification learning. *Hum Brain Mapp* 27:306–313. [CrossRef Medline](#)
- Schönberg T, Daw ND, Joel D, O’Doherty JP (2007) Reinforcement learning signals in the human striatum distinguish learners from nonlearners during reward-based decision making. *J Neurosci* 27:12860–12867. [CrossRef Medline](#)
- Schultz W, Dayan P, Montague PR (1997) A neural substrate of prediction and reward. *Science* 275:1593–1599. [CrossRef Medline](#)
- Seymour B, Daw ND, Roiser JP, Dayan P, Dolan R (2012) Serotonin selectively modulates reward value in human decision-making. *J Neurosci* 32:5833–5842. [CrossRef Medline](#)
- Shepard RN (1987) Toward a universal law of generalization for psychological science. *Science* 237:1317–1323. [CrossRef Medline](#)
- Shepard RN, Hovland CI, Jenkins HM (1961) Learning and memorization of classifications. *Psychol Monogr* 75:1–42.
- Shohamy D, Myers CE, Onlaor S, Gluck MA (2004) Role of the basal ganglia in category learning: how do patients with parkinson’s disease learn? *Behav Neurosci* 118:676–686. [CrossRef Medline](#)
- Smith JD, Minda JP, Washburn DA (2004) Category learning in rhesus monkeys: a study of the Shepard, Hovland, and Jenkins (1961) tasks. *J Exp Psychol Gen* 133:398–414. [CrossRef Medline](#)
- Soto FA, Gershman SJ, Niv Y (2014) Explaining compound generalization in associative and causal learning through rational principles of dimensional generalization. *Psychol Rev* 121:526–558. [CrossRef Medline](#)
- Sutton RS, Barto AG (1998) *Reinforcement learning: an introduction*. Cambridge, MA: MIT.
- Trabasso T, Bower G (1964) Presolution reversal and dimensional shifts in concept identification. *J Exp Psychol* 67:398–399. [CrossRef Medline](#)
- van Spaendonck KP, Berger HJ, Horstink MW, Borm GF, Cools AR (1995) Card sorting performance in Parkinson’s Disease: a comparison between acquisition and shifting performance. *J Clin Exp Neuropsychol* 17:918–925. [CrossRef Medline](#)
- Wei P, Müller HJ, Pollmann S, Zhou X (2011) Neural correlates of binding features within or cross-dimensions in visual conjunction search: an fMRI study. *Neuroimage* 57:235–241. [CrossRef Medline](#)
- Wickens JR, Reynolds JN, Hyland BI (2003) Neural mechanisms of reward-related motor learning. *Curr Opin Neurobiol* 13:685–690. [CrossRef Medline](#)
- Wickens JR, Kötter R (1995) Cellular models of reinforcement. In: *Models of information processing in the basal ganglia* (Houk JC, Davis JL, Beiser DG, eds), pp 187–214. Cambridge, MA: MIT.
- Wilson RC, Niv Y (2011) Inferring relevance in a changing world. *Front Hum Neurosci* 5:189. [CrossRef Medline](#)
- Zhou H, Desimone R (2011) Feature-based attention in the frontal eye field and area v4 during visual search. *Neuron* 70:1205–1217. [CrossRef Medline](#)