Article

Neuron

Dynamic Interaction between Reinforcement Learning and Attention in Multidimensional Environments

Highlights

- Attention constrains reinforcement learning processes to relevant task dimensions
- Learned values of stimulus features drive shifts in the focus of attention
- Attention biases value and reward prediction error signals in the brain
- Dynamic control of attention is associated with activity in frontoparietal network

Authors

Yuan Chang Leong, Angela Radulescu, Reka Daniel, Vivian DeWoskin, Yael Niv

Correspondence

ycleong@stanford.edu (Y.C.L.), angelar@princeton.edu (A.R.), yael@princeton.edu (Y.N.)

In Brief

Leong, Radulescu et al. used eye tracking and fMRI to empirically measure fluctuations of attention in a multidimensional decision-making task. The authors demonstrate that decision making in multidimensional environments is facilitated by a bidirectional interaction between attention and trial-and-error reinforcement learning processes.

Leong et al., 2017, Neuron *93*, 451–463
January 18, 2017 © 2017 Elsevier Inc.
http://dx.doi.org/10.1016/j.neuron.2016.12.040

CellPress

Neuron
Article

Dynamic Interaction between Reinforcement Learning and Attention in Multidimensional Environments

Yuan Chang Leong,[1,6,*] Angela Radulescu,[2,6,*] Reka Daniel,[3] Vivian DeWoskin,[4] and Yael Niv[2,5,7,*]
[1]Department of Psychology, Stanford University, Stanford, CA 94305, USA
[2]Department of Psychology, Princeton University, Princeton, NJ 08544, USA
[3]Dstillery, New York, NY 10016, USA
[4]Trinity Partners, San Francisco, CA 94111, USA
[5]Princeton Neuroscience Institute, Princeton University, Princeton, NJ 08544, USA
[6]Co-first author
[7]Lead Contact
*Correspondence: ycleong@stanford.edu (Y.C.L.), angelar@princeton.edu (A.R.), yael@princeton.edu (Y.N.)
http://dx.doi.org/10.1016/j.neuron.2016.12.040

SUMMARY

Little is known about the relationship between attention and learning during decision making. Using eye tracking and multivariate pattern analysis of fMRI data, we measured participants' dimensional attention as they performed a trial-and-error learning task in which only one of three stimulus dimensions was relevant for reward at any given time. Analysis of participants' choices revealed that attention biased both value computation during choice and value update during learning. Value signals in the ventromedial prefrontal cortex and prediction errors in the striatum were similarly biased by attention. In turn, participants' focus of attention was dynamically modulated by ongoing learning. Attentional switches across dimensions correlated with activity in a frontoparietal attention network, which showed enhanced connectivity with the ventromedial prefrontal cortex between switches. Our results suggest a bidirectional interaction between attention and learning: attention constrains learning to relevant dimensions of the environment, while we learn what to attend to via trial and error.

INTRODUCTION

The framework of reinforcement learning (RL) has been instrumental in advancing our understanding of the neurobiology of trial-and-error learning and decision making (Lee et al., 2012). Yet, despite the widespread success of RL algorithms in explaining behavior and neural activity on simple learning tasks, these same algorithms become notoriously inefficient as the number of dimensions in the environment increases (Bellman, 1957; Sutton and Barto, 1998). Nevertheless, animals and humans faced with high-dimensional learning problems on a daily basis seem to solve them with ease.

How do we learn efficiently in complex environments? One possibility is to employ selective attention to narrow down the dimensionality of the task (Jones and Cañas, 2010; Niv et al., 2015; Wilson and Niv, 2012). Selective attention prioritizes a subset of environmental dimensions for learning while generalizing over others, thereby reducing the number of different states or stimulus configurations that the agent must consider. However, attention must be directed toward dimensions of the environment that are important for the task at hand (i.e., dimensions that predict reward) to provide learning processes with a suitable state representation (Gershman and Niv, 2010; Wilson et al., 2014). What dimensions are relevant to any particular task is not always known and might itself be learned through experience. In other words, for attention to facilitate learning, we might first have to learn what to attend to. We therefore hypothesize that a bidirectional interaction exists between attention and learning in high-dimensional environments.

To test this, we had human participants perform a RL task with compound stimuli—each comprised of a face, a landmark, and a tool—while we scanned their brain using fMRI. At any one time, only one of the three stimulus dimensions was relevant to predicting reward, mimicking real-world learning problems where only a subset of dimensions in the environment is relevant for the task at hand. Using eye tracking and multivariate pattern analysis (MVPA) of fMRI data, we obtained a quantitative measure of participants' attention to different stimulus dimensions on each trial. We then used trial-by-trial choice data to test whether attention biased participants' valuation of stimuli, their learning from prediction errors, or both processes. We generated estimates of participants' choice value and outcome-related prediction errors using the best-fitting model and regressed these against brain data to further determine the influence of attention on neural value and prediction error signals. Finally, we analyzed trial-by-trial changes in the focus of attention to study how attention was modulated by ongoing experience and to search for neural areas involved in the control of attention.

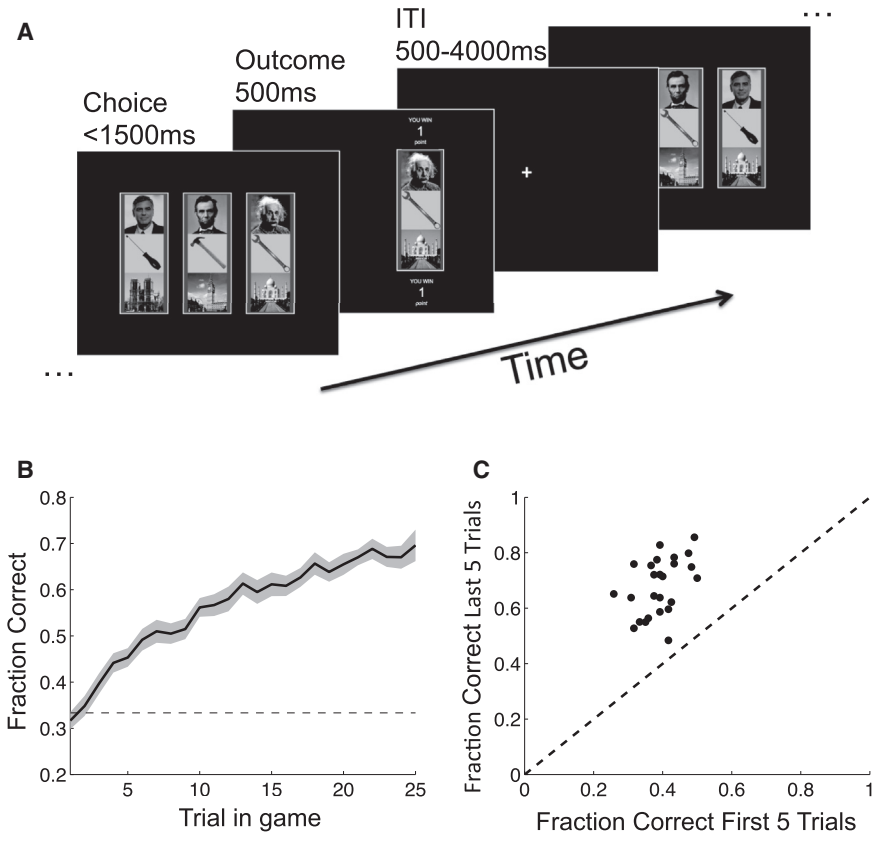

Figure 1. The Dimensions Task

(A) Schematic illustration of the task. On each trial, the participant was presented with three stimuli, each defined along face, landmark, and tool dimensions. The participant chose one of the stimuli, received feedback, and continued to the next trial.

(B) The fraction of trials on which participants chose the stimulus containing the target feature (i.e., the most rewarding feature) increased throughout games. Dashed line: random choice, shading: SEM.

(C) Participants (dots) chose the stimulus that included the target feature (correct stimulus) significantly more often in the last five trials of a game as compared to the first five trials (t_{24} = 15.42, p < 0.001). Diagonal: equality line.

RESULTS

25 participants performed a learning task with multidimensional stimuli and probabilistic reward ("Dimensions Task," Figure 1A). On each trial, participants chose one of three compound stimuli—each comprised of a face, a landmark, and a tool—receiving point rewards. In any one game, only one stimulus dimension (e.g., tools) was relevant for predicting reward. Within that dimension, one "target feature" (e.g., wrench) was associated with a high probability of reward (p = 0.75) and other features were associated with a low probability of reward (p = 0.25). The relevant dimension and target feature changed every 25 trials, and this change was signaled to participants ("New game starting"). Through trial and error, participants learned to choose the stimulus containing the target feature over the course of a game (Figures 1B and 1C). We defined a learned game as one in which participants chose the target feature on every one of the last five trials. By this metric, participants learned on average 11.3 (SE = 0.7) out of 25 games. The number of learned games did not depend on the relevant dimension (F(2,24) = 0.886, p = 0.42).

Attention-Modulated Reinforcement Learning

We obtained two quantitative trial-by-trial measures of participants' attention to each dimension. First, using eye tracking, we computed the proportion of time participants looked at each dimension on each trial. Second, using MVPA, we quantified face-, landmark-, and tool-selective neural activity on each trial (Norman et al., 2006; see Experimental Procedures). Each measure provided a vector of three "attention weights" per trial, denoting the proportion of attention toward each of the dimensions on that trial. Average attention weights were similar across dimensions (Figures S1 and S2), indicating that neither measure was biased toward a particular dimension. The two measures were only moderately correlated (r = 0.34, SE = 0.03), suggesting that the separate measures were not redundant. We therefore used their smoothed product as a composite measure of attention on each trial, which we incorporated into different RL models (see Experimental Procedures).

Attention can modulate RL in two ways. First, attention can bias *choice* by differentially weighing features in different dimensions when computing the value of a composite, multidimensional stimulus. Second, attention can bias *learning* such that the values of features on attended dimensions are updated more as a result of a prediction error. To test whether and how attention modulated learning in our task, we fit four different RL models to the trial-by-trial choice data. In all four models, we assumed that participants chose between the available stimuli based on their expected value, computed as a linear combination of "feature values" associated with the three features of each stimulus, and that feature values were updated after every trial using a prediction error signal (Figure 2) (Rescorla and Wagner, 1972; Sutton and Barto, 1998). In the "uniform attention" (UA) model, all dimensions were weighted equally when computing and updating values. That is, the value of a stimulus was the average values of all its features, and once the outcome of the choice was revealed, the prediction error was equally divided among all features of the chosen stimulus. In the "attention at choice" (AC) model, the value of each stimulus was computed as a weighted sum of feature values, with attention to the respective dimensions on that trial serving as weights. All dimensions were still weighted equally at learning. In contrast,

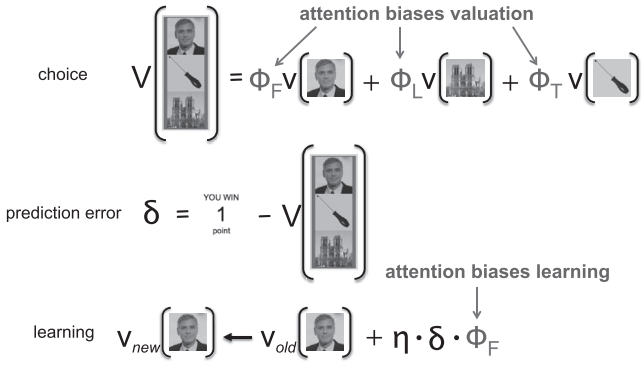

Figure 2. Schematic of the Learning Models

At the time of choice, the value of a stimulus (V) is a linear combination of its feature values (v), weighted by attention weights of the corresponding dimension (ϕ_F, ϕ_L, ϕ_T = attention weights to faces, landmarks, and tools, respectively). In the UA and AL models, attention weights for choice were one-third for all dimensions. Following feedback, a prediction error (δ) is calculated. The prediction error is then used to update the values of the three features of the chosen stimulus (illustrated here only for the face feature), weighted by attention to that dimension and scaled by the learning rate (η). In the UA and AC models, all attention weights were one-third at learning.

in the ''attention at learning'' (AL) model, the update of feature values was differentially weighted by attention; however, all dimensions were equally weighted when computing the value of stimuli for choice. Finally, the ''attention at choice and learning'' (ACL) model combined the AC and AL models so that both choice and learning were biased by attention. The free parameters of each model were fit to each participant's choices (see Experimental Procedures; Table S1).

Both Choice and Learning Are Biased by Attention

To assess how well the models explained participants' behavior, we used a leave-one-game-out cross-validation procedure to compute the likelihood per choice for each participant (Experimental Procedures). The higher the average choice likelihood, the better the model predicted the data of that participant. As a second metric for model comparison, we computed the Bayesian information criterion (BIC; Schwarz, 1978) for each model. In all cases, using the composite attention measure provided a better fit to the data than if we used the MVPA measure or the eye-tracking measure alone (Figure S4); hence, we use the composite attention measure in subsequent analyses unless otherwise noted (see Figures S1–S3 for characterization of the three attention measures).

Both model-performance metrics showed that the ACL model, in which attention modulated both choice and learning, outperformed the other three models (Figure 3). Average likelihood per trial for the ACL model was highest for 21 of 25 subjects, on average significantly higher than that for the AC (t_{24} = 4.72, $p < 0.001$), AL (t_{24} = 6.70, $p < 0.001$), and UA (t_{24} = 8.61, $p < 0.001$) models (Figure 3A). Both the AC and AL models also yielded significantly higher average likelihood per trial than the UA model (AC: t_{24} = 8.03, $p < 0.001$; AL: t_{24} = 7.20, $p < 0.001$). Model comparison using BIC confirmed that the ACL model has the lowest (i.e., best) BIC score (Figure 3B). Moreover, the average likelihood per trial of the ACL model diverged significantly from that of the other models early in the game, when performance was still well below asymptote (as early as trial 2 for AL and UA and from trial 7 for AC; Figure 3C). These results were not driven by the learned portion of games (in which participants may have focused solely on the relevant dimension), as they held when tested on unlearned games only (Figure S5).

The ACL model used the same set of attention weights for choice and learning; however, previous theoretical and empirical work suggest that attention at choice might focus on stimuli or features that are most predictive of reward (Mackintosh, 1975), whereas at learning, one might focus on features for which there is highest uncertainty (Pearce and Hall, 1980). To test whether attention at learning and attention at choice were separable, we took advantage of the higher temporal resolution of the eye-tracking measure. We considered eye positions from 200 ms after stimulus onset to choice as indicating ''attention at choice'' and eye positions during the 500 ms of outcome presentation as a measurement of ''attention at learning.'' Attention at choice and attention at learning on the same trial were moderately correlated (average r = 0.56), becoming increasingly correlated over the course of a game ($F(24,24) = 4.95$, $p < 0.001$; Figure S6). This suggests that as participants figured out the relevant dimension, they attended to the same dimension in both phases of the trial. When we fit the ACL model using attention at choice to bias value computation and attention at learning to bias value update, the model performed slightly, but significantly, better than the ACL model that used whole-trial attention weights for both choice and learning (Figure S7). This suggests that attentional processes at choice and at learning may reflect dissociable contributions to decision making (see also Supplemental Experimental Procedures).

Overall, these results suggest that attention processes biased both how values were computed during choice and how values were updated during learning. Notably, the partial attention models (AC and AL) also explained participants' behavior better than the model that assumed uniform attention across dimensions (UA), providing additional support for the role of attention in participants' learning and decision-making processes.

Neural Value Signals and Reward Prediction Errors Are Biased by Attention

Having found behavioral evidence that attention biased both choice and learning, we hypothesized that neural computations would exhibit similar biases. Previous work has identified two neural signals important for RL processes—an expected value signal in the ventromedial prefrontal cortex (vmPFC; e.g., Hare et al., 2011; Krajbich et al., 2010; Lim et al., 2011) and a reward prediction error signal in the striatum (e.g., O'Doherty et al., 2004; Seymour et al., 2004). Our four models made different assumptions about how attention biases choice and learning and, as such, generated different estimates of expected value and prediction error on each trial.

To test which model was most consistent with the neural value representation, we entered the trial-by-trial value estimates of the chosen stimulus generated by all four models into a single GLM (GLM1 in Experimental Procedures). This allowed us to search the whole brain for clusters of brain activity whose

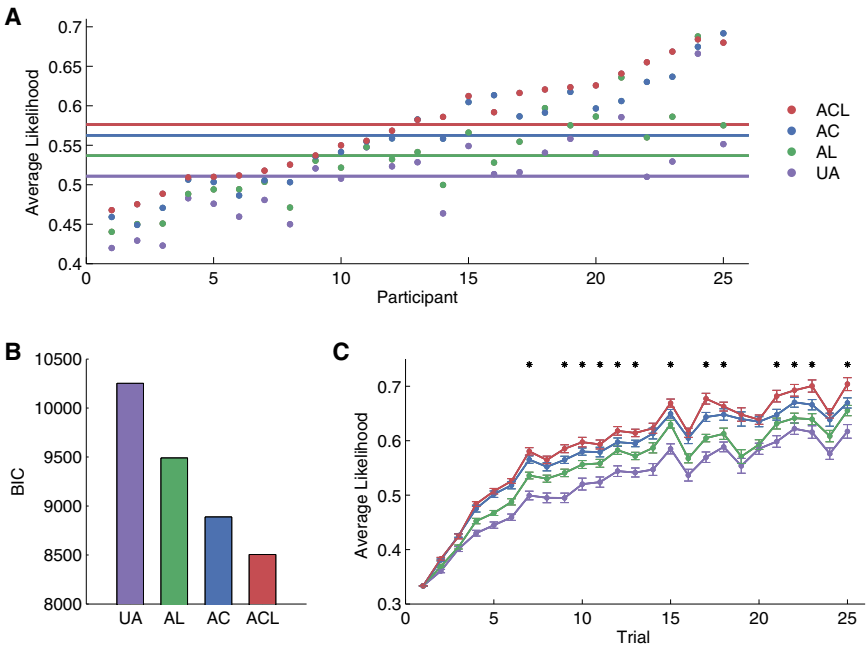

Figure 3. The "Attention at Choice and Learning" Model Explains Participants' Choice Data Best

(A) Average choice likelihood per trial for each model and each participant (ordered by likelihood of the model that best explained their data) shows that the Attention at Choice and Learning (ACL) model predicted the data significantly better than other models (paired t tests, $p < 0.001$). This suggests that attention biased both choice and learning in our task. Solid lines: mean for each model across all participants.

(B) BIC scores aggregated over all participants also support the ACL model (lower scores indicate better fits to data).

(C) Average choice likelihood of the ACL model was significantly higher than that of the AL and UA models from the second trial of a game and onward (paired t tests, $p < 0.05$) and was higher than that of the AC model from as early as trial 7 (paired t tests against AC model, *$p < 0.05$). By the end of the game, the ACL model could predict choice with ~70% accuracy. Error bars, within-subject SE.

variance was uniquely explained by one of the models while simultaneously controlling for the value estimates of the other models. Results showed that activity in the vmPFC was significantly correlated with the value estimates of the ACL model (Figure 4A), suggesting that the computation and update of the value representation in vmPFC was biased by attention. No clusters were significantly correlated with the value estimates of the AC and AL model; one cluster in the visual cortex was significantly correlated with the value estimates of the UA model (Table S2).

Next, we investigated whether neural prediction error signals were also biased by attention. For this, we entered trial-by-trial prediction error regressors generated by each of the four models into a single whole-brain GLM (GLM2 in Experimental Procedures). Prediction errors generated by the ACL model were significantly correlated with activity in the striatum (Figure 4B; Table S2), the area most commonly associated with prediction error signals in fMRI studies. Prediction error estimates of the other models were not significantly correlated with any cluster in the brain. Together, these results provide neural evidence that attention biases both the computation of subjective value as well as the prediction errors that drive the updating of those values.

Attention Is Modulated by Value and Reward

In our previous analyses, we demonstrated that attention biased both choice and learning. In the subsequent analyses, we focus on the other side of the bidirectional relationship, examining how learning modulates attention. As a measure of participants' attention bias (i.e., how strongly participants were attending to one dimension rather than the other two), we computed the standard deviation (SD) of the three attention weights on each trial. Low SD corresponded to relatively uniform attention, while high SD implied that attention was more strongly directed to a subset of dimensions.

Participants' attention bias increased over the course of a game (linear mixed effects model, main effect of trial: $t(24) = 2.79$, $p = 0.01$), and the increase was marginally greater for learned games than for unlearned games (trial number × learned games interaction: $t(22.5) = 1.8$, $p = 0.08$, Figure S3A), as expected from a narrowing of attention when participants learn the target feature and relevant dimension. In parallel, the correlation between the vectors of attention weights for consecutive trials increased steadily over the course of a game, consistent with attention becoming increasingly consistent across trials (main effect of trial number: $t(24.0) = 5.6$, $p < 0.001$, Figure S3B). This effect was also more pronounced for learned games than for unlearned games (trial × learned games interaction: $t(36.5) = 2.96$, $p = 0.005$).

Next, we asked whether learned value (i.e., expected reward) modulated *where* participants directed their attention. We found that the most strongly attended dimension was often also the dimension with the feature of highest value ($M = 0.61$, $SE = 0.02$; significantly higher than chance, $p < 0.001$ bootstrap test, see Supplemental Experimental Procedures), suggesting that attention was often directed toward aspects of the stimuli that had acquired high value. We then tested the interaction between expected reward and the *strength* of the attention bias. A tercile split on the trials based on the strength of attention bias revealed that the proportion of trials in which attention was directed to the dimension with the highest feature value was higher when attention bias was strong (first tercile, $M = 0.72$, $SE = 0.03$) than when attention bias was moderate (second tercile, $M = 0.61$, $SE = 0.02$) and weak (third tercile, $M = 0.49$, $SE = 0.02$) ($p < 0.001$ for all differences, Figure 5A).

We also predicted that feature values would modulate attention such that the greater the difference in feature values across dimensions, the greater the attention bias toward the feature of the highest value. To test this prediction, we performed a tercile

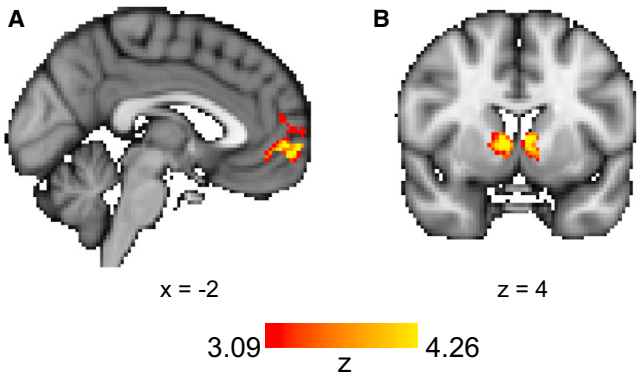

Figure 4. Neural Value and Reward Prediction Error Signals Are Biased Both by Attention for Choice and by Attention for Learning

(A) BOLD activity in the vmPFC was significantly correlated with the value estimates of the ACL model, controlling for the value estimates of the AC, AL, and UA models.

(B) BOLD activity in the striatum correlated with reward prediction errors generated by the ACL model, controlling for reward prediction error estimates from the AC, AL, and UA models. In sum, the ACL model's predictions for value and prediction errors best corresponded to their respective neural correlates.

split of each participant's data based on the standard deviation of the values (SDV) of the highest-valued feature in each dimension, a measure of the difference in feature values across dimensions. Attention bias was stronger on high SDV trials than on middle ($p = 0.030$) and low ($p = 0.027$) SDV trials (Figure 5B) and marginally higher for the middle SDV trials compared to the low SDV trials ($p = 0.067$). Taken together, these results suggest that learned values modulated where participants attended to and the strength of their attention bias.

Building on these results, we hypothesized that attention switches would be more likely when feature values were similar across dimensions. We defined an attention switch as a change in the maximally attended dimension. Our attention measurements suggested that participants switched their focus of attention on approximately one-third of the trials (M = 8.28 [SE = 0.4] switch trials per 25-trial game). Switches were more frequent on low SDV trials than on middle ($p = 0.014$) and high ($p = 0.004$) SDV trials. Switches were also more frequent on middle SDV trials than on high SDV trials ($p = 0.002$) (Figure 5C). To determine the influence of recent reward on attention switches, we ran a logistic regression predicting attention switches from outcome (reward versus no reward) on the preceding five trials. We found that absence of reward in previous trials (up to four trials back) was a significant predictor of attention switches on the current trial t ($p < 0.01$), with regression coefficients decreasing and no longer significantly different from zero at trial t-5 ($t_{24} = 1.72$, $p = 0.09$; Figure 5D). In other words, participants were more likely to switch their focus of attention after a string of no reward, and feedback on more recent trials had a greater influence than trials further back in the past.

Finally, we complemented these analyses with a model-based analysis of attention. Here, we compared different models of the trial-by-trial dynamic allocation of attention (see Experimental Procedures). In particular, we tested whether attention allocation could be better explained by choice history (i.e., attention was enhanced for features that have been previously chosen), reward history (i.e., attention was enhanced for features that had been previously rewarded), or learned value (i.e., attention was enhanced for features associated with higher value over the course of a game). Cross-validated model comparison revealed that the empirical attention data were best explained by a model that tracked feature values. In particular, the "Value" model outperformed the next-best "Recent Reward History" model for both the eye-tracking (lowest root-mean-square deviation [RMSD] in 17/25 subjects, paired-sample t test, $t(24) = 2.77$, $p < 0.05$, Figure 5E) and composite attention (lowest RMSD in 18/25 subjects, paired-sample t test, $t(24) = 2.41$, $p < 0.05$, Figure S8) measures. For the MVPA data, the Value model did not significantly improve upon the predictions of the Recent Reward History model (lowest RMSD in 16/25 subjects, paired-sample t test, $t(24) = 1.02$, $p = 0.31$, Figure 5F); however, it still performed significantly better than the "Recent Choice History" model (paired-sample t test, $t(24) = 3.83$, $p < 0.001$).

In summary, both model-based and model-free results suggest that attention was dynamically modulated by ongoing learning. As participants learned to associate value with features over the course of a game, attention was directed toward dimensions with features that acquired high value (which, in our task, are also the features that are most predictive of reward) in accord with Mackintosh's theory of attention (Mackintosh, 1975). The greater the feature values in a dimension, the stronger the attention bias was toward that dimension. Conversely, when feature values across dimensions were similar, attention was less focused and switches between dimensions were more likely. Finally, attention was better explained as a function of learned value rather than simpler models of reward or choice history.

Attention Switches Correlate with Activity in a Frontoparietal Control Network

Our results suggest that ongoing learning and feedback dynamically modulated participants' deployment of attention. How might the brain be realizing these attention dynamics? To answer this, we searched for brain areas that were more active during switches in attention. As in our previous analyses, we labeled trials on which the maximally attended dimension was different from that of the preceding trial as switch trials and the rest as stay trials. A contrast searching for more activity on switch rather than stay trials (GLM3 in Experimental Procedures) showed clusters in the dorsolateral prefrontal cortex (dlPFC), intraparietal sulcus (IPS), frontal eye fields (FEF), pre-supplementary motor area (preSMA), precuneus, and fusiform gyrus (Figure 6; Table S3). These brain regions are part of a frontoparietal network that has been implicated in the executive control of attention (Corbetta and Shulman, 2002; Petersen and Posner, 2012). Our results suggest that this attentional-control system also supports top-down allocation of attention during learning and decision making in multidimensional environments.

We next asked whether this network was activated only by attention switches, or perhaps it was involved in the

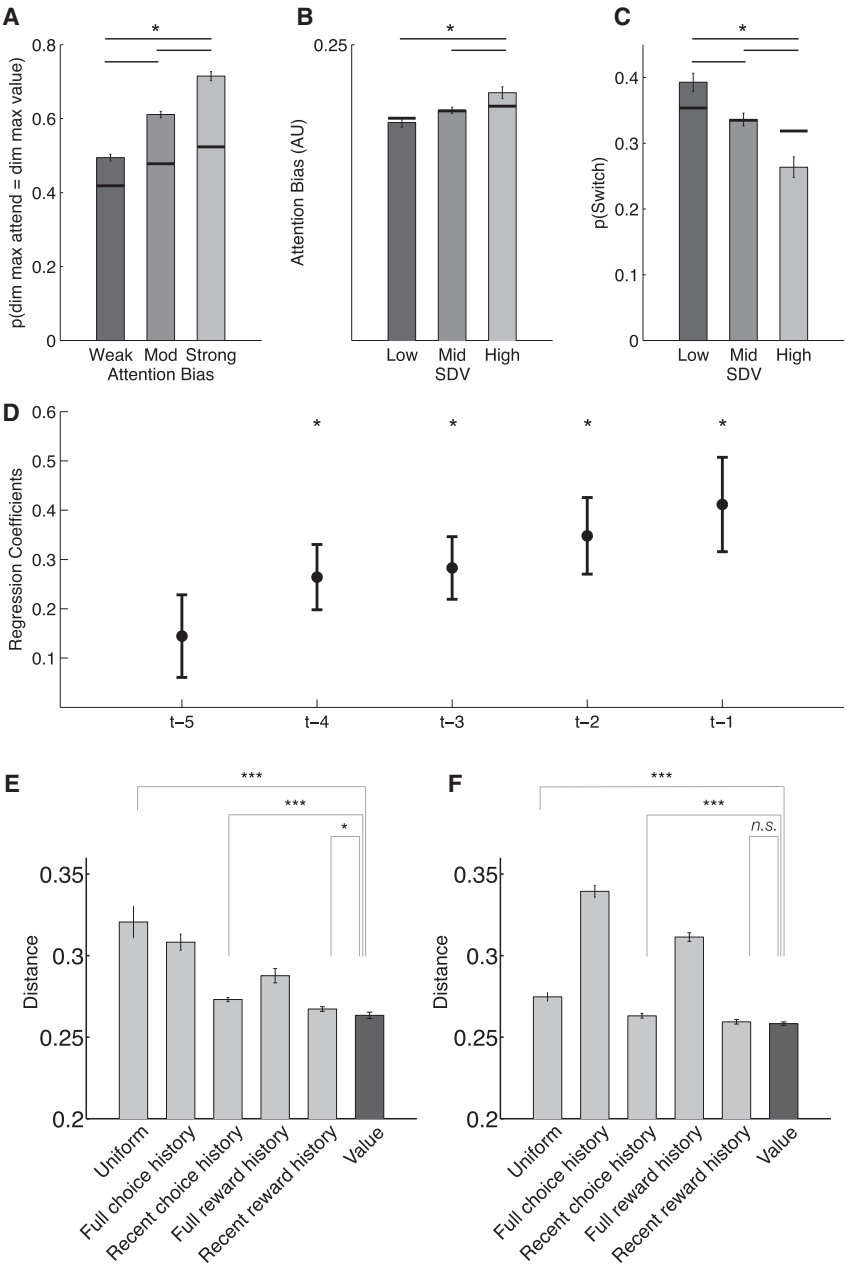

Figure 5. Attention Is Modulated by Ongoing Learning

(A) Proportion of trials in which the most attended dimension was the dimension that included the highest-valued feature. This proportion, in all cases significantly above chance, was highest in trials with strong attention bias as compared to those with moderate and weak attention bias.

(B) Trials with a higher SD of the highest values across dimensions (SDV) showed stronger attention bias.

(C) The probability of an attention switch was highest on low SDV trials. Overall, the greater the difference between feature values, the stronger the attention bias and the less likely participants were to switch attention. Black lines: means of corresponding null distributions generated from a bootstrap procedure in which attention weights for each game were replaced by weights from a randomly selected game from the same participant.

(D) Coefficients of a logistic regression predicting attention switches from absence of reward on the preceding trials. Outcomes of the past four trials predicted attention switches reliably.

(E and F) Comparison of models of attention fitted separately to the eye-tracking (E) and the MVPA (F) measures, according to the root-mean-square deviation (RMSD) of the model's predictions from the empirical data (lower values indicate a better model). Plotted is the subject-wise average per-trial RMSD calculated from holdout games in leave-one-game-out cross-validation (Experimental Procedures). The Value model (dark grey) has the lowest RMSD. Error bars, 1 SEM. ***p < 0.001; **p < 0.01; *p < 0.05.

accumulation of evidence leading up to an attention switch. The latter hypothesis would predict that activity in these regions would ramp up on trials prior to a switch. To test for this, we defined regions of interest (ROIs) in the dlPFC, IPS, and preSMA using Neurosynth (http://neurosynth.org). For each ROI, we extracted the mean time course during each run and modeled these data using a GLM with regressors for attention switch trials, as well as four trials preceding each switch (GLM4 in Experimental Procedures). In all three ROIs, we found that activity increased only on attention-switch trials and not the trials preceding them, suggesting that this network was involved in switching attention rather than accumulating evidence for the switch (Figure 7).

Enhanced Functional Connectivity between vmPFC and Frontoparietal Network between Switches

Finally, we searched for the neural mechanism mediating the interaction between learning and attention. For this, we performed a psychophysiological interaction (PPI) analysis that searched for brain areas that exhibit enhanced functional connectivity with the vmPFC that is specific to switch trials or to stay trials (GLM5 in Experimental Procedures). This showed that connectivity between parts of the frontoparietal attention network and vmPFC differed according to whether participants switched attention on the current trial or not: bilateral dlPFC and preSMA (as well as the striatum and ventrolateral PFC) were significantly anti-correlated with vmPFC on stay trials, above and beyond the baseline functional connectivity between these regions (Figure 8; Table S4). This suggests that as value—signaled by the vmPFC—increased, activity in the dlPFC and preSMA decreased, reducing the tendency to switch attention between task dimensions (and vice versa when the value signal decreased). The connectivity on switch trials was not significantly different from baseline. However, we note that interpretations of the direction of interaction are difficult as they are relative to the baseline connectivity between vmPFC and other regions

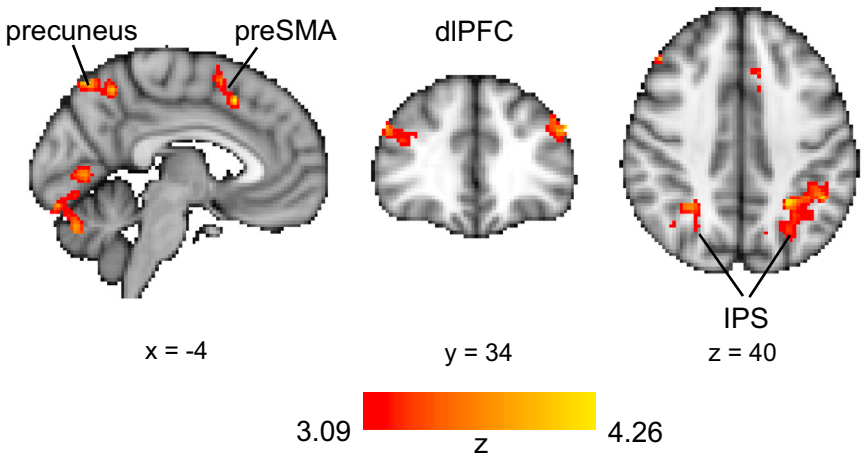

Figure 6. Neural Correlates of Attention Switches
BOLD activity in a frontoparietal network was higher on switch trials than on stay trials.

(as modeled in the GLM), and therefore the above interpretation should be treated with caution.

DISCUSSION

Learning and attention play complementary roles in facilitating adaptive decision making (Niv et al., 2015; Wilson and Niv, 2012). Yet, there has been surprisingly little work in cognitive neuroscience addressing how attention and learning interact. Here, we combined computational modeling, eye tracking, and fMRI to study the interaction between trial-and-error learning and attention in a decision-making task. We used eye tracking and pattern classification of multivariate fMRI data to measure participants' focus of attention as they learned which of three dimensions of task stimuli was instrumental to predicting and obtaining reward. Model-based analysis of both choice and neural data indicated that attention biased how participants computed the values of stimuli and how they updated these values when obtained reward deviated from expectations. The strength and focus of the attention bias was, in turn, dynamically modulated by ongoing learning, with trial-by-trial allocation of attention best explained as following learned value rather than the history of reward or choices. Blood-oxygen-level-dependent (BOLD) activity in a frontoparietal executive control network correlated with switches in attention, suggesting that this network is involved in the control of attention during reinforcement learning.

Our study builds on a growing body of literature in which RL models are applied to behavioral and neural data. Converging evidence suggests that the firing of midbrain dopamine neurons during reward-driven learning corresponds to a prediction error signal that is key to learning (Lee et al., 2012; Schultz et al., 1997; Steinberg et al., 2013). The dopamine prediction error hypothesis has generated much excitement, as it suggests that RL algorithms provide a formal description of the mechanisms underlying learning. However, it is becoming increasingly apparent that this story is far from complete (Dayan and Niv, 2008; O'Doherty, 2012). In particular, RL algorithms suffer from the "curse of dimensionality": they are notoriously inefficient in realistic, high-dimensional environments (Bellman, 1957; Sutton and Barto, 1998).

How can the RL framework be extended to provide a more complete account of real-world learning? A key insight is that learning can be facilitated by taking advantage of regularities in tasks. For example, humans can aggregate temporally extended actions into subroutines that reduce the number of decision points for which policies have to be learned (Botvinick, 2012). Here, we highlight a parallel strategy whereby participants employ selective attention to simplify the state representation of the task. While real-world decisions often involve multidimensional options, not all dimensions are relevant to the task at hand. By attending to only the task-relevant dimensions, one can effectively reduce the number of environmental states to learn about. In our task, for example, attending to only the face dimension simplifies the learning problem to one with three states (each of the faces) rather than 27 states corresponding to all possible stimulus configurations. Selective attention thus performs a similar function as dimensionality-reduction algorithms that are often applied to solve computationally complex problems in the fields of machine learning and artificial intelligence (Ponsen et al., 2010).

Drawing on theories of visuospatial attention (Desimone and Duncan, 1995), we conceptualized attention as weights that determine how processing resources are allocated to different aspects of the environment. In our computational models of choice behavior, these weights influenced value computation in choice and value update in learning. Several previous studies have taken a similar approach to investigate the relationship between attention and learning (Jones and Cañas, 2010; Marković et al., 2015; Wilson and Niv, 2012; Wunderlich et al., 2011), and recently, we demonstrated that neural regions involved in control of attention are also engaged during learning in multidimensional environments, providing neural evidence for the role of attention in learning (Niv et al., 2015). These prior studies, however, have relied on inferring attention weights indirectly from choice behavior or from self-report.

Here, we obtained a direct measure of attention, independent of choice behavior, using eye tracking and MVPA analysis of fMRI data. Attention and eye movements are functionally related (Kowler et al., 1995; Smith et al., 2004) and share underlying neural mechanisms (Corbetta et al., 1998; Moore and Fallah, 2001). Attention is also known to enhance the neural representation of the attended object category (O'Craven et al., 1999), which can be decoded from fMRI data using pattern classification (Norman et al., 2006). Therefore, as a second proxy for attention, we quantified the level of category-selective neural patterns of activity on each trial. By incorporating attention weights derived from the two measures into computational models fitted to

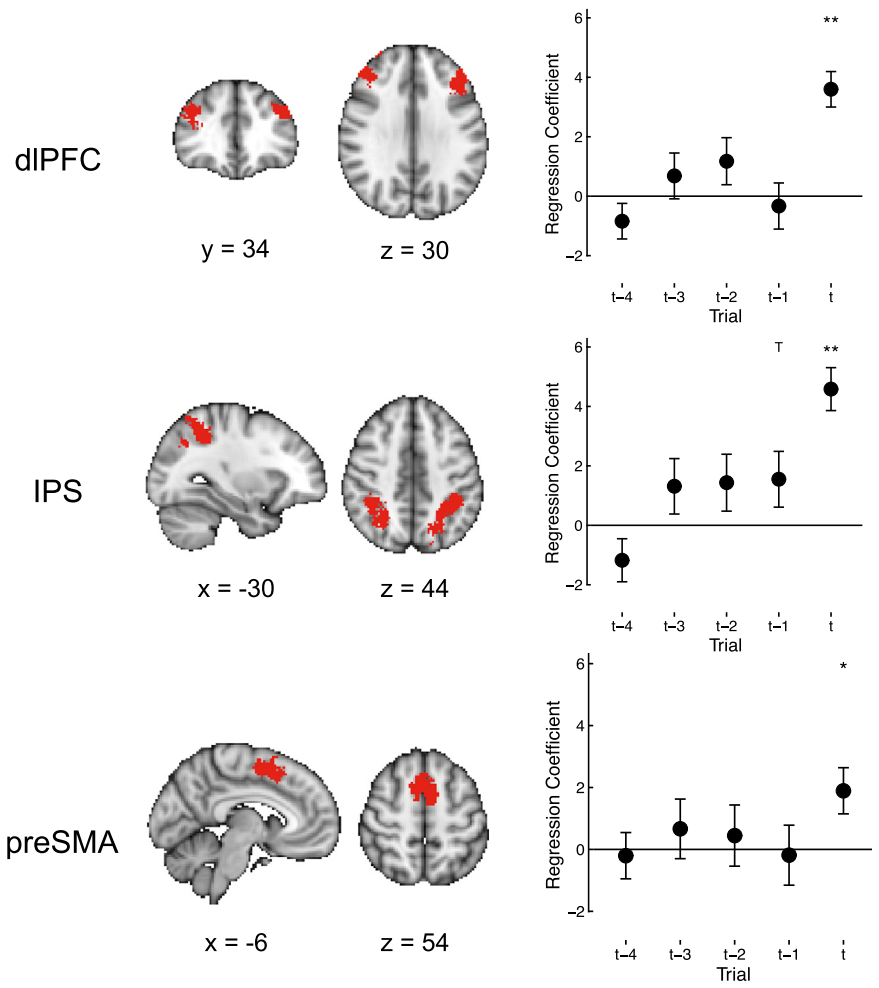

Figure 7. The Frontoparietal Attention Network Is Selectively Activated on Switch Trials

Left: ROIs in the frontoparietal attention network defined using Neurosynth. Right: regression coefficients for switch trials (t) and four trials preceding a switch (t-1 to t-4). Mean ROI activity increased at switch trials (regression weights for trial t are significantly positive), but not trials preceding a switch. Error bars, SEM. $^{**}p < 0.01$, $^{*}p < 0.05$, $^{T}p < 0.1$.

participants' choices, we provide evidence for the influence of attention processes on both value computation and value updating during RL.

Previous work has shown that value computation is guided by attention (Krajbich et al., 2010) and that value signals in the vmPFC are biased by attention at the time of choice (Hare et al., 2011; Lim et al., 2011). For example, Hare et al. (2011) found that when attention was called to the health aspects of food choices, value signals in the vmPFC were more responsive to the healthiness of food options, and participants were more likely to make healthy choices. Here, we extend those findings and demonstrate that attention biases not only value computation during choice, but also the update of those values following feedback. Another neural signal guiding decision making is the reward prediction error signal, which is reflected in BOLD activity in the striatum, a major site of efferent dopaminergic connections (O'Doherty et al., 2004; Seymour et al., 2004). We found that this prediction error signal was also biased by attention, providing additional evidence that RL signals in the brain are attentionally filtered.

But how does the brain know what to attend to? To facilitate choice and learning, attention has to be directed toward stimulus dimensions that are relevant to obtaining reward, such that learning processes operate on the correct state representation of the task. However, at the beginning of each game in our task, participants did not know which dimension is relevant. Our results suggest that without explicit cues, participants can learn to attend to the dimension that best predicts reward and dynamically modulate both what they attend to and how strongly they attend based on ongoing feedback. These findings are consistent with the view of attention as an information-seeking mechanism that selects information that best informs behavior (Gottlieb, 2012). In particular, a model in which attention was allocated based on learned value provided the best fit to the empirical attention measures. Notably, this model is closely related to Mackintosh (1975)'s model of associative learning, which assumes that attention is directed to features that are most predictive of reward.

An alternative view of how attention changes with learning was suggested by Pearce and Hall (1980). According to their model, attention should be directed to the most uncertain features in the environment—that is, the features that participants know the least about and that have been associated with more prediction errors. In support of this theory, errors in prediction have been shown to enhance attention to a stimulus and increase the learning rate for that stimulus (Esber et al., 2012; Holland and Gallagher, 2006). The seemingly contradictory Mackintosh and Pearce-Hall theories of attention have both received extensive empirical support (Pearce and Mackintosh, 2010). Dayan et al. (2000) offered a resolution by suggesting that when making choices, one should attend to the most reward-predictive features, whereas when learning from prediction errors, one should attend to the most uncertain features. When we separately assessed attention at choice and attention at learning in our task, we found that the two measures were correlated. Nevertheless, a model with separate attention weights at choice and learning fit participants' data better than the same model that used the same whole-trial attention weights at both phases. Our results thus support a dissociation between attention at choice and

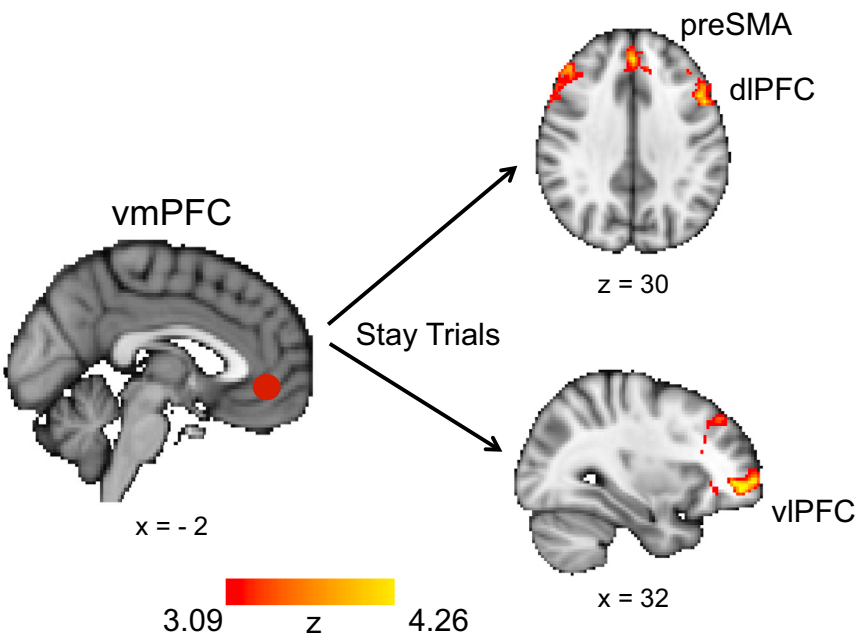

Figure 8. vmPFC and Frontal Attention Regions Were Anticorrelated on Stay Trials

A PPI analysis with vmPFC activity as seed regressor and two task regressors (and PPI regressors) for switch trials and for stay trials revealed that activity in the dlPFC, preSMA, vlPFC, and striatum (not shown) was anticorrelated with activity in the vmPFC specifically on stay trials, above and beyond the baseline connectivity between these areas.

learning, although further work is clearly warranted to determine how attention in each phase is determined. In particular, our task was not optimally designed for separately measuring attention at choice and attention at learning as the outcome was only presented for 500 ms, during which participants also had to saccade to the outcome. Our task is also not well suited to test the Pearce-Hall framework for attention at learning because, in our task, the features associated with more prediction errors are features in the irrelevant dimensions that participants were explicitly instructed to try to ignore.

Neurally, our results suggest that the flexible deployment of attention during learning and decision making is controlled by a frontoparietal network that includes the IPS, FEF, precuneus, dlPFC, and preSMA. This network has been implicated in the executive control of attention in a variety of cognitive tasks (Corbetta and Shulman, 2002; Petersen and Posner, 2012), and the dlPFC in particular is thought to be involved in switching among "task sets" by inhibiting irrelevant task representations when task demands change (Dias et al., 1996; Hyafil et al., 2009). Our findings demonstrate that the same neural mechanisms involved in making cued attention switches can also be triggered in response to internal signals that result from learning from feedback over time. We interpret these findings as suggesting that the frontoparietal executive control network flexibly adjusts the focus of attention in response to ongoing feedback, such that learning can operate on the correct task representation in multidimensional environments.

How does the frontoparietal network know when to initiate an attention switch? Behavioral data suggested that outcomes on preceding trials influenced participants' decision to switch their focus of attention. BOLD activity in this network, however, was only higher on the trial of the switch and not on the preceding trials, suggesting that the frontoparietal network was involved in mediating attention switches rather than accumulating evidence for the switch. Given our finding that learned value influenced the focus and magnitude of attention bias, we also tested for a functional interaction between value representations in the vmPFC and attentional switches. We found evidence of increased anti-correlation between vmPFC and a subset of areas in the frontoparietal network on trials in which attention was not switched, supporting a role for high learned values in decreasing the tendency to switch attention, and of low values in instigating attentional switches. However, we note that the temporal resolution of the BOLD signal prevents us from making strong claims about the directionality of information flow.

In summary, our study provides behavioral and neural evidence for a dynamic relationship between attention and learning—attention biases what we learn about, but we also learn what to attend to. By incorporating attention into the reinforcement learning framework, we provide a solution for the seemingly computationally intractable task of learning and decision making in high-dimensional environments. Our study also demonstrates the potential of using eye tracking and MVPA to measure trial-by-trial attention in cognitive tasks. Combining such measures of attention with computational modeling of behavior and neural data will be useful in future studies of how attention interacts with other cognitive processes to facilitate adaptive behavior.

EXPERIMENTAL PROCEDURES

Participants

29 participants were recruited from the Princeton community (10 male, 19 female, ages 18–31, mean age = 21.1). Participants were right handed and provided written, informed consent. Experimental procedures were approved by the Princeton University Institutional Review Board. Participants received $40 for their time and a performance bonus of up to $8. Data from four participants were discarded because of excessive head motion (>3 mm) (one participant) or because the participant fell asleep (three participants), yielding an effective sample size of 25 participants.

Stimuli

Stimuli were comprised of nine gray-scale images consisting of three famous faces, three famous landmarks, and three common tools. Stimuli were presented using MATLAB software (MathWorks) and the Psychophysics Toolbox (Brainard, 1997). The display was projected to a translucent screen, which participants could view through a mirror attached to the head coil.

Experimental Task

On each trial, participants were presented with three compound stimuli, each defined by a feature on each of the three dimensions (faces, landmarks, and tools), vertically arranged into a column. Row positions for the dimensions were fixed for each participant and counterbalanced across participants. Stimuli were generated by randomly assigning a feature (without replacement) on each dimension to the corresponding row of each stimulus. Participants had 1.5 s to choose one of the stimuli, after which the outcome was presented for 0.5 s. If participants did not respond within 1.5 s, the trial timed out. The inter-trial interval (ITI) was 2 s, 4 s, or 6 s (truncated geometric distribution, mean = 3.5 s), during which a fixation cross was presented. Stimulus presentations were timelocked to the beginning of a repetition time (TR). In any one game, only one dimension was relevant for predicting reward. Within that dimension, one target feature predicted reward with high probability. If participants chose the stimulus containing the target feature, they had a 0.75 probability of receiving reward. If they chose otherwise, they had a 0.25 probability of receiving reward. The relevant dimension and target feature were randomly determined for each game. Participants were told when a new game started but were not told which dimension was relevant or which feature was the target feature. Participants performed four functional runs of the task, each consisting of six games of 25 trials each.

Localizer Task

We used a modified one-back task to identify patterns of fMRI activation in the ventral visual stream that were associated with attention to faces, landmarks, or tools. Participants observed a display of the nine images similar to that used for the main task. On each trial, they had to attend to one particular dimension. Participants were instructed to respond with a button press if the horizontal order of the three images in the attended dimension repeated between consecutive trials. The order of the images on each trial was pseudo-randomly assigned so that participants would respond on average every three trials. Participants were told which dimension to attend to at the start of each run, and the attended dimension changed every one to five trials, signaled by a red horizontal box around the new attended dimension. The sequence of attended dimensions was counterbalanced (Latin square design) to minimize order effects. On each trial, the stimulus display was presented for 1.4 s, during which participants could make their response. Participants received feedback (500 ms) for hits, misses, and false alarms, but not for correct rejections (where a fixation cross was presented for 500 ms instead). Each trial was followed by a variable ITI (2 s, 4 s, or 6 s, truncated geometric distribution, mean = 3.51 s). Participants performed two runs of the localizer task (135 trials each) after completing the main task.

Eye Tracking

Eye-tracking data were acquired using an iView X MRI-LR system (SMI SensoMotoric Instruments) with a sampling rate of 60 Hz. System output files were analyzed using in-house MATLAB code (see "Measures of Attention" below).

fMRI Data Acquisition and Preprocessing

MRI data were collected using a 3T MRI scanner (Siemens Skyra). Anatomical images were acquired at the beginning of the session (T1-weighted MPRAGE, TR = 2.3 s, echo time [TE] = 3.1 s, flip angle = 9°, voxel size 1 mm^3). Functional images were acquired in interleaved order using a T2*-weighted echo planar imaging (EPI) pulse sequence (34 transverse slices, TR = 2 s, TE = 30 ms, flip angle = 71°, voxel size 3 mm^3). Image volumes were preprocessed using FSL/FEAT v.5.98 (FMRIB software library, FMRIB). Preprocessing included motion correction, slice-timing correction, and removal of low-frequency drifts using a temporal high-pass filter (100 ms cutoff). For MVPA analyses, we trained and tested our classifier in each participant's native space. For all other analyses, functional volumes were first registered to participants' anatomical image (rigid-body transformation with 6° of freedom) and then to a template brain in Montreal Neurological Institute (MNI) space (affine transformation with 12° of freedom).

Measures of Attention

Eye Tracking

A horizontal rectangular area of interest (AOI) was defined around each horizontal dimension in the visual display. Data were preprocessed by low-pass filtering (10 Hz cutoff) to reduce high-frequency noise, discarding data from the first 200 ms after the onset of each trial to account for saccade latency and taking the proportion of time participants' point of gaze resided within each AOI as a measure of attention to the corresponding dimension. The level of noise in the eye-tracking measure can vary systematically between participants. To account for subject-specific noise, we computed a weighted sum between the raw measure and uniform attention (one-third to each dimension). The weight ω_{ET}, which served to smoothly interpolate between uniform attention and the empirical eye-tracking measure, was a free parameter fit to each subject's behavioral data. As ω_{ET} decreased, the empirical measure contributed less to the final attention vector. Fitting a subject-specific ω_{ET} parameter provided us with a data-driven method to weigh the empirical measure based on how much it contributed to explaining choices. For model comparison, the parameter was fit using leave-one-game-out cross-validation to avoid over-fitting. For the fMRI analyses, the parameter was fit to all games (see "Choice Models" below).

MVPA

A linear support vector machine (SVM) was trained on data from the localizer task to classify the dimension that participants were attending to on each trial based on patterns of BOLD activity. Analysis was restricted to voxels in a ventral visual stream mask consisting of the bilateral occipital lobe and ventral temporal cortex. The mask was created in MNI space using anatomical masks defined by the Harvard-Oxford Cortical Structural Atlas as implemented in FSL. The mask was then transformed into each participant's native space using FSL's FLIRT implementation, and classification was performed in participants' native space. Cross-validation classification accuracy on the localizer task was 87.4% (SE = 0.9%; chance level: 33%). The SVM was then applied to data from the Dimensions Task to classify participants' trial-by-trial attention to the three dimensions. Classification was performed using the SVM routine LinearNuSVMC (Nu = 0.5) implemented in the PyMVPA package (Hanke et al., 2009; see also Supplemental Experimental Procedures). On each trial, the classifier returned the probability that the participant was attending to each of the dimensions (three numbers summing to 1). Similar to the eye-tracking measure of attention, we computed a weighted sum between the probabilities and uniformly distribution attention, where the weight ω_{MVPA} was a free parameter fit to each participant's behavioral data.

Composite Measure

To combine the two measures of attention, we computed a composite measure as the product of eye-tracking and MVPA measures of attention, renormalized to sum to 1. Taking a product means that each of the two measures contributes to the composite according to how strongly the measure is biased toward one dimension and not others. For example, a uniform (1/3, 1/3, 1/3) measure contributes nothing to the composite measure for that trial. In contrast, if one measure is extremely biased to one dimension (e.g., 1, 0, 0), it overrides the other measure completely.

Behavioral Performance

Trials were scored as correct if the participant chose the stimulus containing the target feature of that game. We computed individual learning curves by averaging the number of correct trials in each trial position across all games. We then computed the group learning curve by averaging the individual learning curves over all participants. Overall performance was assessed with a paired t test that tested whether the fraction of correct trials was significantly higher on the last five trials than on the first five trials. We defined a learned game as a game in which the participant chose the stimulus containing the target feature on each of the last five trials of the game. A one-way repeated-measures ANOVA was used to test whether learning a game depended on the relevant dimension of that game.

Choice Models

We tested four RL models (Sutton and Barto, 1998). All four models assumed that participants learned to associate each feature with a value and linearly combined the values of features to obtain the value of a compound stimulus:

$$V_{(t)}(S_i) = \sum_d \phi_t(d) \cdot v_t(d, S_i) \qquad \text{Equation 1}$$

where $V_t(S_i)$ is the value of stimulus i on trial t, $\phi_t(d)$ is the attention weight of dimension d and $v_t(d, S_i)$ denotes the value of the feature in dimension d of stimulus S_i. Following feedback, a prediction error, δ_t, was calculated as the difference between observed reward, r_t, and the expected value of the chosen stimulus $V_t(S_c)$:

$$\delta_t = r_t - V_t(S_c) \quad \text{Equation 2}$$

δ_t was then used to update the feature values of the chosen stimulus

$$v_{t+1}(d, S_c) = v_t(d, S_c) + \eta \cdot \phi_t(d) \cdot \delta_t \quad \text{Equation 3}$$

where the update is weighted by attention to the respective dimensions and scaled by a learning rate or step-size parameter η, which was fit to each participants' behavioral data. Because we computed one prediction error and one update per trial, this model is an instance of Rescorla and Wagner (1972)'s learning rule.

In the ACL model, both value computation and value update were biased by attention weights. In the AC model, the attention measure was used for value computation, but all $\phi_t(d)$ were set to one-third during value update such that the three dimensions were updated equally during learning. In the AL model, the attention measure was used for value update, but all $\phi_t(d)$ were set to one-third for value computation, weighting all dimensions equally at choice. In the UA model, $\phi_t(d)$ were set to one-third for both value computation and value update.

For all models, choice probabilities were computed according to a softmax action selection rule:

$$\pi_t(c) = \frac{e^{\beta V_t(S_c)}}{\sum_a e^{\beta V_t(S_a)}} \quad \text{Equation 4}$$

where $\pi_t(c)$ is the probability of choosing stimulus c, a enumerates over the three available stimuli, and β is a free inverse-temperature parameter that determines how strongly choice is biased toward the maximal-valued stimulus.

The three attention models (ACL, AL, and AC) had four free parameters—ω_{MVPA}, ω_{ET}, β, and η—while the uniform attention model had two free parameters, β and η. Model comparison used a leave-one-game-out cross-validation procedure: for each participant and for each game, we fit the model to participants' choices from all other games by minimizing the negative log likelihood of the choices. Given these parameters, we then calculated the likelihood of each choice in the held-out game. The total likelihood of the data of each participant, computed for each game as it was held out, was then divided by the number of trials that the participant played to obtain the geometric average of the likelihood per trial. Because we used cross-validation, we could compare between models based on their likelihood per trial without fear of over-fitting and did not need to correct for model complexity. Best-fit model parameters (fitted to all games to maximize power) are reported in Table S1. As a second metric, we compared the models using the BIC (Schwarz, 1978; see Supplemental Experimental Procedures).

Modulation of Attention

We analyzed the attention weights used in the ACL model and investigated how they changed over time, as well as how they were modulated by value and reward (Figures S1–S3). These analyses are described in detail in the Supplemental Experimental Procedures.

Attention Models

We developed a series of computational models that made predictions about the allocation of attention to face, landmark, and tool dimensions on each trial. The "Full Choice History" model allocated attention based on a leaky choice count. On each trial, counts for each of the three chosen features were incremented by 1, and counts for the remaining six unchosen features were decayed toward 0 at a subject-specific decay rate. Attention to each dimension was then determined by the softmax of the maximal count on each dimension. That is, on each trial, we took the highest count among the three features of each dimension and passed them through a softmax function (see Equation 4) to obtain three attention weights that sum up to 1.

The "Recent Choice History" model used a delta-rule update to adjust the weights of the chosen features toward 1. For each chosen feature, the weight $w_t(d, S_{chosen})$ was updated as $w_{t+1}(d, S_{chosen}) = w_t(d, S_{chosen}) + \eta_a[1 - w_t(d, S_{chosen})]$, where η_a is a free update rate parameter. Here, too, the weights of unchosen features were decayed toward 0 at a subject-specific decay rate, and the predicted attention weights were determined using softmax on the maximum weights in each dimension.

The "Full Reward History" model allocated attention based on a leaky *reward* count: on rewarded trials, counts of chosen features were incremented by 1 and counts of unchosen features were decayed toward 0 at a subject-specific decay rate. No learning or decay occurred on unrewarded trials. Again, softmax was applied to the maximum counts in each dimension to determine attention.

In the "Recent Reward History" model, analogous to the Recent Choice History model, on each rewarded trial a delta-rule update adjusted the weights of the chosen features toward 1, and weights of the unchosen features were decayed toward 0.

Finally, in the "Value" model, attention tracked feature values. The value of each stimulus was assumed to be the sum of the values of all its features. Feature values were initialized at 0 and updated via reinforcement learning with decay (see also Niv et al., 2015): on each trial, a prediction error was calculated as the difference between the obtained reward and the value of the chosen stimulus. The value of chosen features was updated based on the prediction error scaled by a subject-specific update rate, while the value of unchosen features was decayed toward 0 at a subject-specific decay rate. As in the other models, the maximum feature value in each dimension was then passed through a softmax function to obtain the predicted attention vector. Unlike the Recent Reward History model, this model learned not only from positive, but also from negative prediction errors and based error-driven learning on a stimulus-level prediction error.

In sum, the Full Choice and Recent Choice models allocated attention based on prior choices, whereas in the Full Reward, Recent Reward, and Value models, the history of reinforcement determined fluctuations in attention. The Full Choice and Full Reward models had two free parameters—decay rate and softmax gain. The Recent Choice, Recent History, and Value models had an additional free parameter—the update rate. We also compared the models to a baseline zero-parameter model in which attention is always uniform (1/3, 1/3, 1/3). As with our models of choice behavior, we evaluated the attention models using leave-one-game-out cross-validation: for each participant and for each game, we fit the free parameters of the model to all but that game by minimizing the RMSD of the predicted attention weights from the measured attention weights. We then used the model to predict attention weights for the left-out game to determine the mean RMSD per trial for each model. We fit the models separately for the raw (preprocessed but not smoothed) eye-tracking attention measure, the raw (unsmoothed) MVPA attention measure, and the composite measure (with smoothing parameters ω_{ET} and ω_{MVPA} determined according to the best fit to choice behavior).

fMRI Analyses

We implemented five linear models (GLMs) as design matrices for analysis of the fMRI data:

GLM1 served to investigate whether the computation and update of the expected value signal in the brain was biased by attention. For each participant, we generated estimates for the expected value of the chosen stimulus on each trial using the UA, AC, AL, and ACL models. We entered these value estimates into the GLM as parametric modulators of the stimulus onset regressor. We did not orthogonalize the regressors because, in linear regression, variance shared by different regressors is automatically not attributed to any of the regressors. GLM1 could therefore identify regions that are associated with the value estimates of each model, while simultaneously controlling for the value estimates of the other models. Reaction time, trial outcome, outcome onset, and head movement parameters were also added as nuisance regressors. With the exception of head movement parameters, all regressors were convolved with the hemodynamic response function. Missed-response trials were not modeled as there was no chosen stimulus in those trials. The GLM

was estimated throughout the whole brain using FSL/FEAT v.5.98 available as part of the FMRIB software library (FMRIB). We imposed a family-wise error cluster-corrected threshold of $p < 0.05$ (FSL FLAME 1), with a cluster-forming threshold of $p < 0.001$. Unless otherwise stated, all GLM analyses included the same nuisance regressors and were corrected for multiple comparisons using the same procedure.

GLM2 served to investigate whether prediction error signals were also biased by attention. This GLM was identical to GLM1 except that instead of generating estimates of expected values, we generated estimates of trial-by-trial prediction errors using the four models and entered them into the GLM as parametric modulators of the outcome onset regressor.

GLM3 modeled switch and stay trials as stick functions at the onset of the respective trials. We defined switch trials as trials in which the maximally attended dimension was different from the previous trial and all the rest as stay trials. A contrast identified clusters that were more active during switch versus stay trials.

GLM4 included one regressor with stick functions at the onset of all switch trials, another with stick functions at the onset of all trials immediately preceding a switch trial, and so forth for the four trials preceding each switch, totaling to five trial-onset regressors. We used this GLM to analyze a set of ROIs in the frontoparietal attention network, dlPFC, IPS, and preSMA, pre-defined using the online meta-analytical tool Neurosynth (http://neurosynth.org). For each region, we generated a meta-analytic reverse inference map (dlpfc: 362 studies; ips: 173 studies; pre-sma: 125 studies). We thresholded each map at $z > 5$ and retained only the cluster containing the peak voxel in each hemisphere. For each ROI, we then extracted the activity time course averaged across voxels. Since we ran the GLM on average ROI activity (one test for each ROI), we report uncorrected p values, though we note that most of the results would survive a Bonferroni corrected threshold of $p < 0.016$ (0.05/3).

GLM5 was used for a PPI analysis to find areas in the brain exhibiting differences in functional connectivity with the vmPFC on switch trials versus stay trials. A separate GLM first identified clusters in the brain that were associated with the value of the chosen stimulus on each trial, as estimated using the ACL model. This map was thresholded at $p < 0.001$ to obtain a group-level vmPFC ROI. We then defined a participant-specific vmPFC ROI by thresholding the participant-level map at $p < 0.01$ and retaining clusters that fell within the group-level ROI. For each participant, we extracted the mean time course from this ROI and used it as the seed regressor for the PPI. We generated two task regressors—one for switch trials, one for stay trials—each modeled as a stick function with value of +1 at stimulus onset and 0 otherwise and convolved with the hemodynamic function. We then generated two PPI regressors by taking the product of each task regressor and the vmPFC time course.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, eight figures, and four tables and can be found with this article online at http://dx.doi.org/10.1016/j.neuron.2016.12.040.

AUTHOR CONTRIBUTIONS

All authors conceived and designed the experiment. Y.C.L. and V.D. piloted the experiment. Y.C.L. and A.R. ran the experiment. Y.C.L. and R.D. analyzed fMRI data. Y.C.L. performed computational modeling of choice data. A.R. performed computational modeling of attention data. Y.C.L., A.R., and Y.N. wrote the paper. Y.N. supervised the project and acquired funding.

ACKNOWLEDGMENTS

We thank Jamil Zaki and Ian Ballard for scientific discussions and helpful comments on earlier versions of the manuscript and members of the Y.N. lab for their comments and support. This work was supported by the Human Frontier Science Program Organization, grant R01MH098861 from the National Institute for Mental Health, and grant W911NF-14-1-0101 from the Army Research Office. The views expressed do not necessarily reflect the opinion or policy of the federal government and no official endorsement should be inferred.

Received: March 18, 2016
Revised: November 3, 2016
Accepted: December 28, 2016
Published: January 18, 2017

REFERENCES

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

Botvinick, M.M. (2012). Hierarchical reinforcement learning and decision making. Curr. Opin. Neurobiol. *22*, 956–962.

Brainard, D.H. (1997). The psychophysics toolbox. Spat. Vis. *10*, 433–436.

Corbetta, M., and Shulman, G.L. (2002). Control of goal-directed and stimulus-driven attention in the brain. Nat. Rev. Neurosci. *3*, 201–215.

Corbetta, M., Akbudak, E., Conturo, T.E., Snyder, A.Z., Ollinger, J.M., Drury, H.A., Linenweber, M.R., Petersen, S.E., Raichle, M.E., Van Essen, D.C., and Shulman, G.L. (1998). A common network of functional areas for attention and eye movements. Neuron *21*, 761–773.

Dayan, P., and Niv, Y. (2008). Reinforcement learning: the good, the bad and the ugly. Curr. Opin. Neurobiol. *18*, 185–196.

Dayan, P., Kakade, S., and Montague, P.R. (2000). Learning and selective attention. Nat. Neurosci. *3*, 1218–1223.

Desimone, R., and Duncan, J. (1995). Neural mechanisms of selective visual attention. Annu. Rev. Neurosci. *18*, 193–222.

Dias, R., Robbins, T.W., and Roberts, A.C. (1996). Dissociation in prefrontal cortex of affective and attentional shifts. Nature *380*, 69–72.

Esber, G.R., Roesch, M.R., Bali, S., Trageser, J., Bissonette, G.B., Puche, A.C., Holland, P.C., and Schoenbaum, G. (2012). Attention-related Pearce-Kaye-Hall signals in basolateral amygdala require the midbrain dopaminergic system. Biol. Psychiatry *72*, 1012–1019.

Gershman, S.J., and Niv, Y. (2010). Learning latent structure: carving nature at its joints. Curr. Opin. Neurobiol. *20*, 251–256.

Gottlieb, J. (2012). Attention, learning, and the value of information. Neuron *76*, 281–295.

Hanke, M., Halchenko, Y.O., Sederberg, P.B., Hanson, S.J., Haxby, J.V., and Pollmann, S. (2009). PyMVPA: a python toolbox for multivariate pattern analysis of fMRI data. Neuroinformatics *7*, 37–53.

Hare, T.A., Malmaud, J., and Rangel, A. (2011). Focusing attention on the health aspects of foods changes value signals in vmPFC and improves dietary choice. J. Neurosci. *31*, 11077–11087.

Holland, P.C., and Gallagher, M. (2006). Different roles for amygdala central nucleus and substantia innominata in the surprise-induced enhancement of learning. J. Neurosci. *26*, 3791–3797.

Hyafil, A., Summerfield, C., and Koechlin, E. (2009). Two mechanisms for task switching in the prefrontal cortex. J. Neurosci. *29*, 5135–5142.

Jones, M., and Cañas, F. (2010). Integrating reinforcement learning with models of representation learning. In Proceedings of the 32nd Annual Conference of the Cognitive Science Society, S. Ohlsson and R. Catrambone, eds. (Cognitive Science Society), pp. 1258–1263.

Kowler, E., Anderson, E., Dosher, B., and Blaser, E. (1995). The role of attention in the programming of saccades. Vision Res. *35*, 1897–1916.

Krajbich, I., Armel, C., and Rangel, A. (2010). Visual fixations and the computation and comparison of value in simple choice. Nat. Neurosci. *13*, 1292–1298.

Lee, D., Seo, H., and Jung, M.W. (2012). Neural basis of reinforcement learning and decision making. Annu. Rev. Neurosci. *35*, 287–308.

Lim, S.-L., O'Doherty, J.P., and Rangel, A. (2011). The decision value computations in the vmPFC and striatum use a relative value code that is guided by visual attention. J. Neurosci. *31*, 13214–13223.

Mackintosh, N.J. (1975). A theory of attention: variations in the associability of stimuli with reinforcement. Psychol. Rev. *82*, 276–298.

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

Moore, T., and Fallah, M. (2001). Control of eye movements and spatial attention. Proc. Natl. Acad. Sci. USA *98*, 1273–1276.

Niv, Y., Daniel, R., Geana, A., Gershman, S.J., Leong, Y.C., Radulescu, A., and Wilson, R.C. (2015). Reinforcement learning in multidimensional environments relies on attention mechanisms. J. Neurosci. *35*, 8145–8157.

Norman, K.A., Polyn, S.M., Detre, G.J., and Haxby, J.V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fMRI data. Trends Cogn. Sci. *10*, 424–430.

O'Craven, K.M., Downing, P.E., and Kanwisher, N. (1999). fMRI evidence for objects as the units of attentional selection. Nature *401*, 584–587.

O'Doherty, J.P. (2012). Beyond simple reinforcement learning: the computational neurobiology of reward-learning and valuation. Eur. J. Neurosci. *35*, 987–990.

O'Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K., and Dolan, R.J. (2004). Dissociable roles of ventral and dorsal striatum in instrumental conditioning. Science *304*, 452–454.

Pearce, J.M., and Hall, G. (1980). A model for Pavlovian learning: variations in the effectiveness of conditioned but not of unconditioned stimuli. Psychol. Rev. *87*, 532–552.

Pearce, J.M., and Mackintosh, N.J. (2010). Two theories of attention: a review and possible integration. In Attention and Associative Learning, C.J. Mitchell and M.E. LePelley, eds. (Oxford University Press), pp. 11–14.

Petersen, S.E., and Posner, M.I. (2012). The attention system of the human brain: 20 years after. Annu. Rev. Neurosci. *35*, 73–89.

Ponsen, M., Taylor, M.E., and Tuyls, K. (2010). Abstraction and generalization in reinforcement learning: a summary and framework. In Adaptive and Learning Agents, M.E. Taylor and K. Tuyls, eds. (Springer Berlin Heidelberg), pp. 1–32.

Rescorla, R.A., and Wagner, A.R. (1972). A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement. In Classical Conditioning II: Current Research and Theory, A.H. Black and W.F. Prokasy, eds. (Appleton-Century-Crofts), pp. 64–99.

Schultz, W., Dayan, P., and Montague, P.R. (1997). A neural substrate of prediction and reward. Science *275*, 1593–1599.

Schwarz, G. (1978). Estimating the dimension of a model. Ann. Stat. *6*, 461–464.

Seymour, B., O'Doherty, J.P., Dayan, P., Koltzenburg, M., Jones, A.K., Dolan, R.J., Friston, K.J., and Frackowiak, R.S. (2004). Temporal difference models describe higher-order learning in humans. Nature *429*, 664–667.

Smith, D.T., Rorden, C., and Jackson, S.R. (2004). Exogenous orienting of attention depends upon the ability to execute eye movements. Curr. Biol. *14*, 792–795.

Steinberg, E.E., Keiflin, R., Boivin, J.R., Witten, I.B., Deisseroth, K., and Janak, P.H. (2013). A causal link between prediction errors, dopamine neurons and learning. Nat. Neurosci. *16*, 966–973.

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

Wilson, R.C., and Niv, Y. (2012). Inferring relevance in a changing world. Front. Hum. Neurosci. *5*, 189.

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

Wunderlich, K., Beierholm, U.R., Bossaerts, P., and O'Doherty, J.P. (2011). The human prefrontal cortex mediates integration of potential causes behind observed outcomes. J. Neurophysiol. *106*, 1558–1569.