Beyond Reward Prediction Errors: Human Striatum Updates Rule Values During Learning

Ian Ballard[1*], Eric M. Miller[2], Steven T. Piantadosi[3], Noah Goodman[2], Samuel M. McClure[4]

1. Stanford Neurosciences Graduate Training Program, Stanford University. Stanford, CA 94305, USA
2. Department of Psychology, Stanford University, Stanford, CA 94305, USA.
3. Department of Brain and Cognitive Sciences, University of Rochester. Rochester, NY 14627, USA.
4. Department of Psychology, Arizona State University, Tempe, AZ 85287, USA.

Corresponding Author
Ian Ballard
Department of Psychology
450 Serra Mall, Stanford CA 94305
iballard@stanford.edu

Conflict of Interest
The authors declare no competing financial interests.

ABSTRACT

Humans naturally group the world into coherent categories defined by membership rules. Rules can be learned implicitly by building stimulus-response associations using reinforcement learning (RL) or by using explicit reasoning. We tested if striatum, in which activation reliably scales with reward prediction error, would track prediction errors in a task that required explicit rule generation. Using functional magnetic resonance imaging during a categorization task, we show that striatal responses to feedback scale with a “surprise” signal derived from a Bayesian rule-learning model. We also find that striatal feedback responses are inconsistent with RL prediction error and demonstrate that striatum and caudal inferior frontal sulcus (cIFS) are involved in updating the likelihood of discriminative rules. We conclude that the striatum, in cooperation with the cIFS, is involved in updating the values assigned to categorization rules, rather than representing reward prediction errors.

1.1 INTRODUCTION

Humans possess a remarkable ability to learn from incomplete information, and rely on multiple strategies to do so. Consider a card game where hearts are rewarded and other cards are not. A simple learning strategy, model-free learning, directly associates stimuli and/or actions with rewards that they predict (Sutton and Barto 1998). This algorithm would efficiently learn that card suits predict different reward values. Now consider a more complex game in which the queen of spades is also rewarded, except when it is paired with all the hearts. A more efficient strategy than model-free learning would be to reason over abstract rules or categories that apply to cards. This strategy requires a cognitive model of the environment based on explicit rules (Miller and Cohen 2001). A large body of work has mapped the neural circuitry underlying model-free learning as well as the circuitry underlying the execution of well-learned cognitive models (Badre and D'Esposito 2009). However, little is known about how cognitive models are acquired or where the variables required to learn models are represented.

Model-free and cognitive model learning have typically been associated with different neural systems: a mesolimbic striatal system for the former and a lateral cortical system for the latter (Glascher et al. 2010; Daw et al. 2011). In model-free learning, striatal neurons encode the value of different stimuli and actions and communicates values to cortical regions via recurrent loops (Haber & Knutson, 2010). Ascending midbrain dopamine projections carry signed reward prediction errors (Montague et al. 1996; Schultz 1997) that underlies the learning of stimulus- and action-outcome associations (Reynolds et al. 2001; Kawagoe et al. 2004; Daw et al. 2005; Morris et al. 2010). Conversely, in cognitive models, prefrontal neuronal pools represent abstract rules, implement their control over behavior, and update rules when appropriate (Buschman et al. 2012). We tested the hypothesis that learning cognitive models depends on parallel neural circuitry to what is known to be involved in reinforcement learning (RL).

This hypothesis may appear straightforward, but it faces several theoretical challenges. First, RL and rule-based learning operate on different information: the former assigns values to stimuli or actions and the latter reasons explicitly over abstract rules, concepts, or structured relationships (Goodman et al. 2008; Glascher et al. 2010; Tenenbaum et al. 2011). We propose that, in addition to encoding stimuli and action values, striatal neurons also encode values of cortical rule representations (e.g., "all hearts and queen of spades"). Second, explicit rule-learning does not require a reward prediction error (RPE). We propose that the mesolimbic dopamine system is not specialized for conveying RPEs; rather, it encodes update signals that reflect new information in a variety of learning contexts.

In order to test this hypothesis, we focused on the robust observation in RL research that striatal activation changes in proportion to RPE (O'Doherty et al. 2003; Rutledge et al. 2010; Garrison et al. 2013). We tested whether the striatum represented RPEs when subjects are biased to learn by reasoning over explicit rules, rather than by the incremental buildup of stimulus-response relationships. If you are learning a card game where hearts are rewarded unless they are paired with a queen, and for many hands you have seen no queens, then discovering that hearts and queens together fail to deliver a reward is highly surprising and also delivers a negative RPE. If the striatum represents RPE as part of a RL algorithm (Garrison et al. 2013), then the striatum should respond negatively to this disappointing outcome. Conversely, if the striatum is involved in updating value representations in response to new information, then it should respond positively to this surprising outcome.

We found that in a task where participants do not use reinforcement learning, striatal feedback response was better described by Bayesian "surprise" than RL prediction error. We reconcile these findings by suggesting that, rather than coding RPE, the striatum updates its value representations in response to new information. Further, we found that posterior inferior frontal sulcus (cIFS), was involved in updating rules and was functionally connected to striatum during feedback. Our findings are consistent with a model in which both rule learning and model-free RL depend on plasticity in cortico-striatal circuits that encode the values of stimuli, actions, and concepts.

MATERIAL & METHODS

2.1 *Participants*

19 participants completed the study (11 female; mean age 21.7 years; SD age 7.3 years). Stanford University's Institutional Review Board approved study procedures, and all participants provided informed consent. Three were excluded because their accuracy was not significantly better than chance. An additional two participants were excluded for excessive head motion (exceeding 2 mm in any direction), leaving 14 participants for analyses.

2.2 *Rule Learning Task*

We used a task that was designed to bias participants towards an explicit rule-learning strategy, rather than relying on incremental learning of stimulus-response contingencies. On each trial, a stimulus was shown that varied on three perceptual dimensions (color: blue or yellow; shape: circle or square; and texture: striped or checkered). Participants assigned stimuli to one of two possible categories, “Dax” or “Bim,” based on perceptual features. Participants were informed that rules linking features to categories changed with each new block of trials. Six different rules were learned in counterbalanced order across blocks: *A* (e.g., blue stimuli are Bims; yellow are Daxes); *A and B* (e.g., blue square stimuli are Bims; all others are Daxes); *A or B*; *(A or B) and C*; *(A and B) or C*; and *A XOR B* (e.g., all blue or square stimuli, with the exception of blue square stimuli are Bims). Feature and category labels that defined each rule were randomly determined at the start of each block. Structurally, the order of trials within a given rule block was identical across participants, enabling direct comparison of performance on a trial-by-trial basis. Participants saw the same order of stimuli (e.g. $\bar{A}B\bar{C}$, $A\bar{B}\bar{C}$, …; where $\bar{A}$ and A refer to the two possible values of a feature) for each rule block, but the mapping of stimulus features (e.g. A to the color blue, or the shape square, etc.) was randomized. For more complex rules, evidence was presented in such a way that a simpler rule sufficed to explain the data for initial trials, and then a discriminating example was presented that required an update to a more complex rule later in the block.

Each trial was divided into three phases: cue, response and feedback. Each phase lasted 2 s and was separated by a random 4 to 6 s delay. During the cue phase, the stimulus to be categorized was shown in the center of the screen. During the response phase, a question mark was shown in the center of the screen, which prompted participants to categorize the stimulus by pressing one of two buttons. During the feedback phase, a message was displayed in the center of the screen that indicated whether the response was “correct” (green text) or “incorrect” (red text).

2.3 *Bayesian rule-learning model*

The rule-learning model was based on the "rational rules" model (Goodman et al. 2008) and implemented in the Python library LOTlib. This model formalizes a statistical learner that operates over the hypothesis space of Boolean propositional logic expressions (e.g. ((A AND B) OR C)). It implicitly defines the infinite hypothesis space of possible expressions using a grammar that permits only valid combinations of logical primitives (AND, OR, NOT) and observable perceptual features. This grammar also defines a prior probability P(H), that biases learners to prefer short expressions (H) that re-use grammar rules. The prior is combined with a likelihood, P(D|H), which quantifies how accurately a rule H predicts observed true/false labels. The likelihood contains a single parameter, α, which corresponds to the probability of generating a random label.

The parameter $\alpha = 0.85$ was fit across all concepts and subjects via grid search. We used Markov-Chain Monte-Carlo to perform inference by sampling hypotheses according to P(H|D). In order to predict incremental responses as the task progressed, sampling was also run incrementally for 10,000 steps at every trial, based on observed data up to that point. The top 100 hypotheses for any amount of data and concept were collected into a set that was treated as a finite hypothesis space for the purposes of efficiently computing model predictions (Piantadosi et al. 2012).

2.4 *RL models*

We designed our task in order to study the mechanism by which humans learn using explicit rules or concepts. Nonetheless, reinforcement learning is a powerful algorithm that can perform well in most tasks, including ours. We sought to buttress our claim that our task involved rule-learning by comparing our Bayesian rule-learning model with simple but powerful RL algorithms that are commonly used in the literature on categorization learning (Niv et al. 2015). We compared three different RL models with different state space representations. Naïve RL had just one feature for each stimulus (blue striped square). Feature RL had a different feature for each stimulus feature (blue), and each stimulus was associated with three features corresponding to each of its features and values were learned over features independently. Ideal RL had a feature associated with each stimulus feature, each pairwise combination of features, and each triplet of features. In this model, each stimulus was associated with seven features: one triplet of stimulus features, three pairwise combinations of features and three individual stimulus features. The value of a state-action pair on trial *t* was determined by taking a weighted sum of each of the feature-action pairs:

$$Q(S, a_i)^t = \sum_{f \in S} W(f, a_i)^t$$

where there are two actions associated with the two choices (Bim or Dax). After trial feedback, the model uses RL to update each feature weight for the next trial:

$$W(f, a_{chosen})^{t+1} = W(f, a_{chosen})^{t} + \alpha[R_t - Q(S, a_{chosen})^{t}]$$

We departed from standard Q modeling by making two opposite updates to the weight of each feature-action pair:

$$W(f, a_{not-chosen})^{t+1} = W(f, a_{not-chosen})^{t} - \alpha[R_t - Q(S, a_{chosen})^{t}]$$

This encodes the symmetry of the task (i.e., if a stimulus is a Bim then it is not a Dax). It also improved the likelihood of the observed data without adding an extra parameter, which only favored RL in model comparison. This change also made the model homologous to a standard value learning model that directly learns the probability that a stimulus is a Bim.

Each models used a softmax decision rule to map task state Q values to the probability of a given action:

$$P(a_i) = \frac{e^{\beta Q(S,a_i)^t}}{\sum_j^2 e^{\beta Q(S,a_j)^t}}$$

Each RL models had two free parameters: α (learning rate) and β (inverse temperature of the softmax).

We fit models by maximizing the likelihood of observed choices for each subject, using Scipy's minimize function with the BGFS method. We constrained the models to fit a single β across subjects and an individual α, to enable a fair comparison with the Bayesian model that fits a single noise parameter across subjects. Fixing β has a regularizing effect on model fitting that biases results away from extreme parameter settings and reduces the correlation between α and β. For neuroimaging analyses, we calculated a single α and β across subjects, which provides additional regularization (Daw et al. 2011).

2.5 Model Comparison

We used the corrected Akaike Information Criterion (AICc) and the Bayesian Information Criterion (BIC) to estimate the posterior model evidence. We used both metrics since they differ in penalization of free parameters (BIC penalizes more strongly). For group model comparison, we used group Bayesian model comparison (Stephan et al. 2009; Rigoux et al. 2014), which is a random effects method for group model comparison robust to outliers.

2.6 *fMRI acquisition*

Functional images were acquired with a 3T General Electric Discovery scanner (Waukesha, WI, USA). Whole-brain BOLD weighted echo planar images were acquired in 40 oblique axial slices parallel to the AC–PC line with a 2000 ms TR (slice thickness = 3.4 mm, no gap, TE = 30 ms, flip angle = 77°, FOV = 21.8 cm, 64 × 64 matrix, interleaved). High-resolution T2-weighted fast spin-echo structural images (BRAVO) were acquired for anatomical reference (TR = 8.2 ms, TE = 3.2 ms, flip angle = 12°, slice thickness = 1.0 mm, FOV = 24 cm, 256 × 256).

2.7 *fMRI analysis*

Preprocessing and whole brain analyses were conducted with Analysis of Functional Neural Images (AFNI; Cox, 1996). Data were slice-time corrected and motion corrected. No participant included in the analyses moved more than 2 mm in any direction. Data used in whole brain analyses were spatially smoothed with a 4 mm FWHM Gaussian filter. Voxel-wise BOLD signals were converted to percent signal change.

We transformed the T2-weighted structural image to Talairach space and applied this transform to preprocessed functional images. Normalized functional images were then analyzed using a general linear model in AFNI. The model contained multiple regressors to estimate responses to each task component, which were then convolved with a two-parameter gamma variate hemodynamic response function. For surface plots, we projected the group data onto the freesurfer template brain using Pysurfer with the default settings and a 6 mm cortical smoothing kernel.

We wanted to identify regions where event-related activity correlated with expected responses related to surprise and updating of rule hypotheses. Parametric regressors derived from the Bayesian model were used to identify brain regions where activation scaled with trial-by-trial estimates of surprise, as well as the degree of change to the hypothesis space as each new exemplar was integrated. This strategy has been recently applied with success in the perceptual domain (O'Reilly et al. 2013). Surprise was calculated as the probability against the label assigned to the stimulus on each trial by the model. Rule updating was calculated as the KL divergence between the posterior distribution on hypotheses before and after the trial.

Three such parametric regressors were included in the model, capturing activation that scaled with 1) surprise at feedback, 2) rule updating at feedback, and 3) rule updating from the previous trial during the subsequent cue period. The final two regressors may seem redundant, but task switching studies show updating occurs when initially possible and at the beginning of subsequent trials, even when inter-trial intervals are several seconds long (Monsell, 2003).

Each event in our general linear model was assumed to occur over a 2 s period (boxcar). We included three additional regressors to model mean activation during the cue, response and feedback periods. A parametric regressor was included to control for activation that varied with reaction time during the response period. Finally, regressors of non-interest were included to account for head movement and third-order polynomial trends in BOLD signal amplitude across the scan blocks.

Maps of t-statistics for each regressor were resampled and transformed into Talairach space. Whole-brain statistical maps were generated using one-sample t-tests at each voxel to localize brain areas with significant loadings on regressors across subjects. Whole-brain maps were thresholded at $p < 0.05$, cluster corrected ($p < 0.005$ voxel-level α with a minimum of 42 contiguous voxels, AlphaSim). Of note, we used AFNI version 16.0.06, in which AlphaSim better estimates type II error. Coordinates are reported in Talairach space with the LAS convention.

A hierarchical analysis was conducted to assess neural responses to Bayesian surprise and RL prediction error. First, activation was modeled with two regressors that encoded activation during positive and negative feedback, as well as task and nuisance variables described above. A second analysis was performed on the residual values obtained from the first regression. Two regressors were included in the second model: a parametric regressor encoding activation that scaled parametrically with surprise during feedback, and a parametric regressor encoding activation that scaled parametrically with prediction error during feedback, with prediction error derived from the best fitting (naïve) RL model. To test for voxels that varied with RL prediction error, we performed a conjunction between the first-level analysis of positive > negative outcomes and the parametric prediction error regressor. To test for voxels that varied with surprise, we performed a conjunction between the first-level analysis of negative > positive outcomes and the parametric surprise regressor. Because prediction errors are strongly correlated with outcome valence, a performed a test for prediction error that accounts for both a main effect of outcome valence and an additional parametric effect once valence has been partialed out. We repeated the above analysis separately for the parametric rule updating analysis to ensure that the cIFS cluster we identified did not merely respond to outcome valence.

We note that outcome valence accounted for 54% of variability in prediction error variance in our best fitting model, and this represents the more important half of variability in prediction error. An RL model that does not have outcome valence information cannot learn to distinguish actions and will respond randomly, while a RL algorithm that has access only to outcome valence, but does not remember recent trials, will implement a win-stay/loose-shift policy. Therefore, the outcome valence analysis is a robust test for prediction error signaling, because a region that does not dissociate outcomes based on valence cannot contribute to a RL algorithm that learns anything about the task.

We conducted region of interest (ROI) analyses in two areas. Based on the results of the rule-updating analyses, in which cIFS correlated with rule-updating and was functionally connected with the striatum, we hypothesized that surprise might be represented in striatal regions that interact with lateral PFC (Haber and Knutson 2009). We used an “executive striatum” ROI taken from a 3-way subdivision of striatum based on diffusion tractography imaging estimated connectivity with cortex (Tziortzi et al.

2013). For the ventral midbrain analysis, we used a probabilistic ventral tegmental area atlas (Murty et al. 2014) based on hand-drawn ROIs (Ballard et al. 2011). We thresholded and binarized the ventral tegmental area atlas at p>0.5.

We also conducted a psychophysical interaction (PPI) in AFNI. First, BOLD signal was averaged over the cluster of significant activation identified in ROI analyses. This time course was mean-centered and included as a regressor in the PPI model. A second regressor was also included, which was the interaction between this time course and feedback. The feedback regressor that was used to make the interaction was centered according to the customary FSL PPI procedure, which ensures the interaction effect is not inflated by correlation that is constant across the timeseries. The PPI model also contained regressors to account for baseline activation during the cue, response, and feedback periods, as well as parametric regressors encoding reaction time and surprise.

1 RESULTS

1.1 *Behavior*

We collected fMRI data while participants completed six 20-trial blocks of a rule-learning task (Figure 1a). In each trial, participants were shown an image and were instructed to classify it as belonging to one of two possible categories (“Dax” or “Bim”) based on perceptual features. Category membership was determined based on logical rules like “stimuli that are either blue or square are Bims” or “stimuli that are both striped and circular are Daxes.” Participants were informed that the rule determining category membership would change at the start of each block. Blocks were separated into clearly demarcated scanning runs in order to minimize interference between rules.

All participants included in analyses performed above chance. Accuracy and reaction times were averaged across participants for each rule block. There was an effect of condition on accuracy (within-subjects ANOVA; $F = 10.1$, $p < 0.0001$) and reaction time ($F=3.1$, $p = 0.03$). Post-hoc t-tests revealed accuracy was lower for complex rules (A XOR B) and ((A and B) or C; Tables S2, S3).

Learning proceeded rapidly, generally reaching an initial asymptote within the first five trials (Figure 1b). For all but the simplest rules, accuracy diminished on later trials, evidenced by accuracy curve “spikes.” These accuracy drops occurred on trials where new, highly informative evidence was presented that required updating from a simple rule to a more complex rule. Accuracy recovered within one or two trials as participants rapidly integrated new evidence.

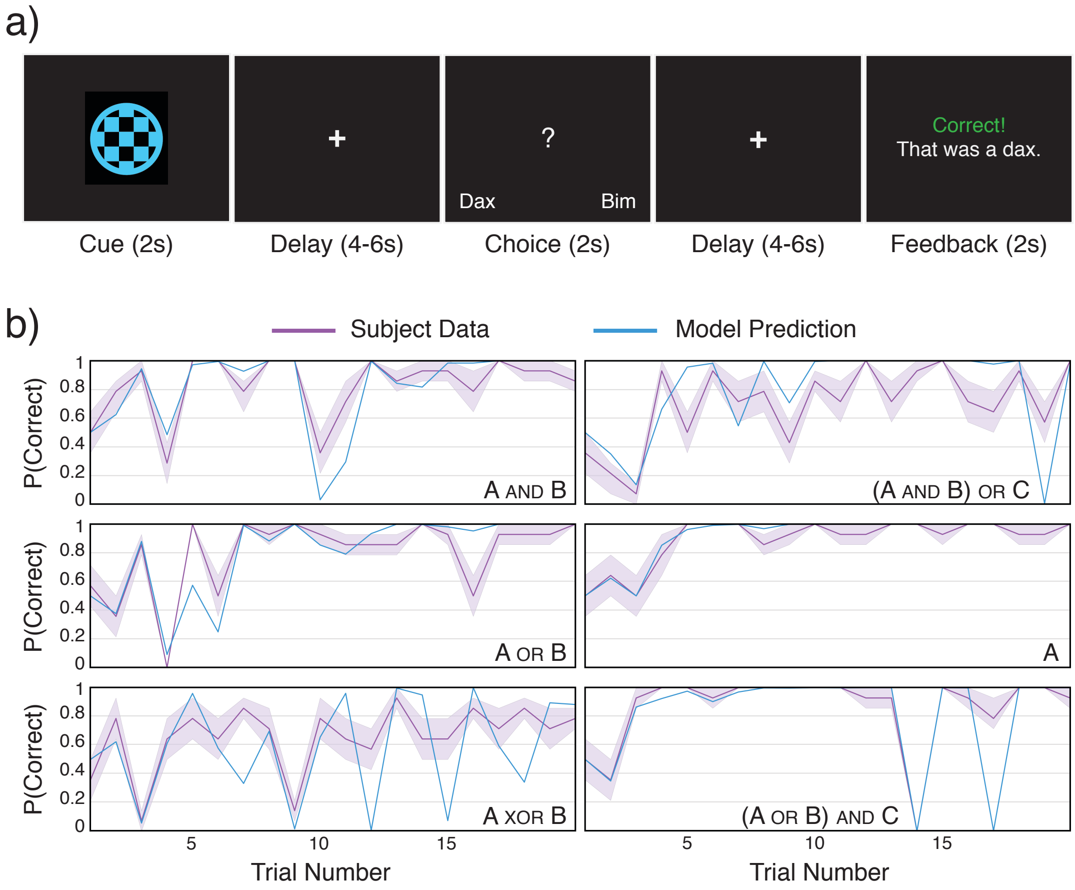

Figure 1: Rule Learning Task and Behavior

a) Participants completed six 20-trial blocks of a rule-learning task. Trials were divided into three phases: cue, response and feedback, each separated by a random 4-6 s delay. During the cue phase (2 s), the stimulus to be categorized was presented in the center of the screen. During the response phase (2 s), a question mark was presented in the center of the screen, prompting participants to press a button to respond. During the feedback phase (2 s), a message was displayed indicating whether the response was correct.

b) Mean participant accuracy and model predictions for each rule. For most rules, accuracy reached an initial asymptote on early trials, but showed decreases on later trials when evidence was introduced that was inconsistent with the simpler hypotheses that sufficed to that point.

1.2 Comparing Bayesian and RL Models

Our task was designed to elicit an explicit learning strategy that relied on the testing of abstract rules. Nonetheless, reinforcement learning is a powerful learning algorithm that can perform well on many tasks, including ours. In addition, there are well-established cortico-striatal circuits for stimulus-response and feature-response learning that could give rise to decent performance on our task (Niv 2009). In order to strengthen our claim that subjects adopted a rule-learning strategy, we used computational modeling of behavior to assess the extent to which subjects employed rule-learning rather than model-free RL. Our Bayesian rule-learning model predicts that subject would use rules that were consistent with previously observed evidence within the block, but with a bias toward parsimony. It also predicts that new evidence will be rapidly integrated as subjects search for a rule that is consistent with all of the observed evidence in a block.

This model predicted behavior across participants, replicating previous work employing similar models (Figure 1b; (Goodman et al. 2008; Piantadosi 2011). Model predictions and average group behavior were correlated (A: $\rho = .97$, $p = 3.8 \times 10^{-12}$; (A and B) or C: $\rho = .72$, $p = 5.6 \times 10^{-4}$; (A or B): $\rho = .83$, $p=8.4 \times 10^{-6}$; (A and B): $\rho = .84$, $p = 3.5 \times 10^{-6}$; (A or B) and C: $\rho = .84$, $p = 3.7 \times 10^{-6}$; (A xor B): $\rho = .6$, $p = 0.005$) and met a Bonferroni corrected significance threshold of $p<0.0083$. We next correlated individual subject behavior and model predictions. We conducted a t-test on Fisher transformed ρ values and observed that the model correlated with individual subject behavior for all rules at a Bonferroni corrected threshold of $p<0.0083$ (A: $\rho = .60$, $p = 2.8 \times 10^{-4}$; (A and B) or C: $\rho = .43$, $p = 1.4 \times 10^{-4}$; (A or B): $\rho = .55$, $p=4.2 \times 10^{-7}$; (A and B): $\rho = .47$, $p = 9.7 \times 10^{-5}$; (A or B) and C: $\rho = .64$, $p = 2.5 \times 10^{-6}$; (A xor B): $\rho = .32$, $p = 9.7 \times 10^{-4}$). We also observed an effect of condition on the correlation between model fit and behavior ($F = 3.28$, $p = 0.04$). It is important to note that these correlations are derived from a model with no free parameters. Consequently, the model predicted rather than fit behavior. In order to compute likelihoods for model comparison, we fit a single noise parameter, corresponding to the probability that participants responded randomly on the given trial, independent of what they had learned up to that point.

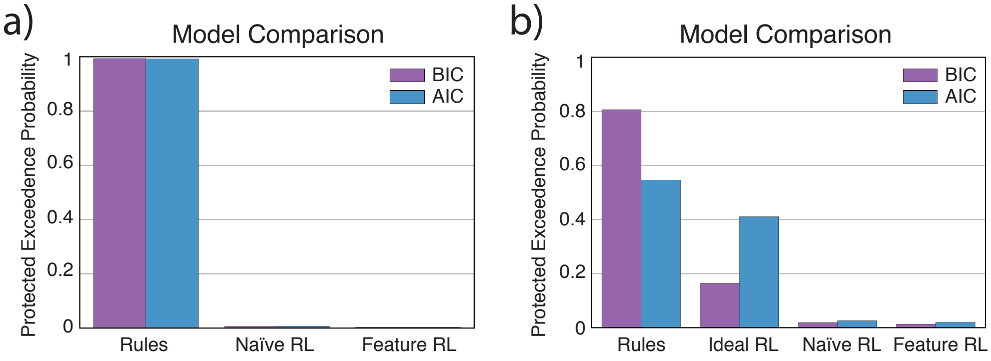

Figure 2: Comparison of Bayesian rule learning and RL

a) Results of a random effects Bayesian model comparison procedure between the Bayesian rule learning model and two standard RL models shows strong evidence for the Bayesian rule model. b) When we include a novel, idealized RL model with a state space designed to perform optimally in the task, the evidence for the Bayesian rule learning model is weaker but still the strongest of all models.

We next compared the Bayesian rule model with simple but powerful RL models from the existing literature on feature learning. We emphasize that humans and animals learn using RL in many circumstances, and the following analysis tests whether our specific task structure was successful in biasing participants towards rule-learning. For example, in feature learning tasks where the feedback is stochastic and only single features (rather than complex conjunctions) are reward predictive, RL provides a better account of behavior than a Bayesian inference model (Niv et al. 2015).

Reinforcement models are a broad class of models and we focused our analysis on models that are 1) commonly used in the literature to describe behavior and 2) are based on the established circuits between sensory cortex and the striatum that support model-free learning (Niv 2009). The first model, naïve RL, learns independently about each stimulus. Since each stimulus repeats 2.5 times, on average, this model can perform quite well. In addition, this model has described behavior well in related tasks (Niv et al. 2015), and captures the essence of stimulus-value coding thought to occur in cortico-striatal pathways (Niv 2009). Because our task is noiseless, it also describes an approach to the task based on episodic memory.

The next model, feature RL, learns about each feature (e.g., blue, circular, striped) independently. This model reflects the idea that the decomposition of stimuli into constituent features in sensory cortex can support cached-value learning based on those features (O'Reilly and Rudy 2001). Feature RL could be expected to learn some rules well, {A, A or B}, but struggle with rules that involve conjunctions of features {A and B}. We fit each RL model to maximize the likelihood of each participant's choices and computed the corrected Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) approximations to the log model evidence (Experimental Procedures). The protected exceedance probability, or the probability that one model is more likely than others above and beyond chance (Stephan et al. 2009), for the Bayesian rule learning model was very high (BIC: 99.5%, AIC: 99.4 %), suggesting that subjects used Bayesian rule learning rather than RL.

These analyses are based on RL models commonly used in the psychological and neuroscience literatures on reinforcement learning and motivated by neural circuits for reinforcement learning. However, we also constructed a RL model that would be ideally suited to learning the rules in our specific task. This third RL model, ideal RL, learns weights for all stimuli (unique combinations of shape, color, texture), each individual feature, and each pairwise combination of features. This optimized RL model can both generalize and learn to represent AND rules.

The ideal RL model is more complex than the naïve RL model. For stimuli with multiple features, the state space grows very rapidly (roughly with n choose 2 for n features), challenging feasibility of this model. Despite these concerns, we fit this model as a best-case scenario of a RL algorithm ideally suited to the demands of our particular task. Model comparison across all four models still favored the Bayesian rule learning model (pEP = 80.6% for BIC and pEP = 54.6% for AIC; Figure 2), but the evidence in favor of the Bayesian rule comparison model was moderate. Rule-by-rule model comparison suggests that the naïve RL model outperforms all of the others on the challenging XOR rule in which the Bayesian rule model does not describe behavior well, as well as (A and (B or C)). The ideal RL model is worse than the Bayesian model for the other complex rules and is a bad model for simple rules. These model comparisons highlight the need for better modeling of XOR learning and show that the Bayesian rule learning model provides a good account of participants' choices even when compared to an RL model using an optimal state space for our task. We conclude that our task manipulation was successful in biasing participants towards an explicit rule-based learning strategy.

1.3 *Striatum activation does not reflect reinforcement learning prediction errors*

Our behavioral analysis suggested that our task design successfully biased participants towards rule-learning rather than stimulus or feature response learning. We next tested whether the striatum reflects prediction errors, which would underlie incremental stimulus response learning, or Bayesian "surprise", which reflects beliefs about abstract rules. The learning signal used by RL algorithms is the prediction error, which is the difference between the predicted strength of the chosen action and an indicator function on whether the action was correct. In our case,

$$PE_t = I(\mathrm{Correct}_t) - Q(S_t, a_{chosen})$$

The analogous signal in our Bayesian rule-learning model, which we call surprise, is equal to the difference between the strength of evidence for a category and the actual category. If on trial *t* a given stimulus $S_t$is a Bim (O_t = Bim), then the surprise is given by

$$Surprise_t = 1 - P(O_t = \ Bim \mid S_1, \dots, S_t;\ O_1, \dots, O_{t-1}\ )$$

Although RL prediction errors and "surprise" are generally correlated, they diverge in an important way. RL prediction error is signed and therefore will almost always be greater for correct than for incorrect responses. Although "unsigned prediction errors" exist in other modeling frameworks such as predictive coding, we emphasize that in any non-degenerate environment, reinforcement learning does not work without a signed prediction error. The signed error is what endows the model with the ability to prefer states and actions that lead to rewards over those that do not. In contrast, surprise will be larger for incorrect than correct outcomes because incorrect predictions are necessarily more surprising. We designed our sequence of stimuli so that initially a simple rule accounted for the data, but later on new data required an update to a more complex rule. This procedure ensured there would be trials with large positive surprise and large negative prediction error.

Surprise was larger for negative outcomes (t = 21.8, p = 3.7 ×10^{-93}, Figure 3b), and outcome valence alone accounted for 22.1% of the variance in surprise. Conversely, RL prediction errors extracted from the best-fitting ideal model were larger for positive outcomes (t = 47.2, p = 2.3 ×10^{-310}, Figure 3a) and outcome valence alone accounted for

57.1% of the variance. The effect of outcome valence on these learning signals was powerful and in opposite directions. These effects are useful because they allow for a simple and non-parametric test of RPE: a region that does not show a larger response to positive than negative outcomes cannot code for a RPE (See Methods for discussion). We therefore predicted that, if the striatum reflects "surprise" in our task, then mean striatal blood oxygen level dependent (BOLD) activation should be larger for negative compared to positive outcomes. This is a novel prediction given the extensive body of work showing larger striatal responses to positive outcomes (Delgado 2007). We emphasize that this prediction is specific to our particular task in which negative feedback is generally more informative than positive feedback.

An analysis of negative versus positive outcomes yielded a large cluster of voxels that encompassed most of the dorsal striatum as well as part of the ventral striatum (Figure 3d). The reverse contrast (positive vs. negative outcome) yielded regions in the dorsomedial prefrontal cortex (dmPFC) and posterior paracentral lobule (pPCL; corrected $p < .05$, Figure 3c), but no striatal regions were found, even at liberal thresholds ($p<0.01$, uncorrected). The striatum is a functionally heterogeneous structure, but this heterogeneity does not follow clear anatomical landmarks (Haber and Knutson 2009). We hypothesized that the striatum was tracking the values of rules maintained in lateral frontal cortex. We therefore examined whether a striatal subregion defined by its connectivity to lateral frontal cortex (Tziortzi et al. 2013), referred to as executive striatum, showed the same BOLD activation pattern. We observed a larger response to negative outcomes in the executive striatum ROI ($t = 3.03$, $p < 0.01$, Figure 4a).

Together, these results are incompatible with RL prediction error signaling in the striatum.

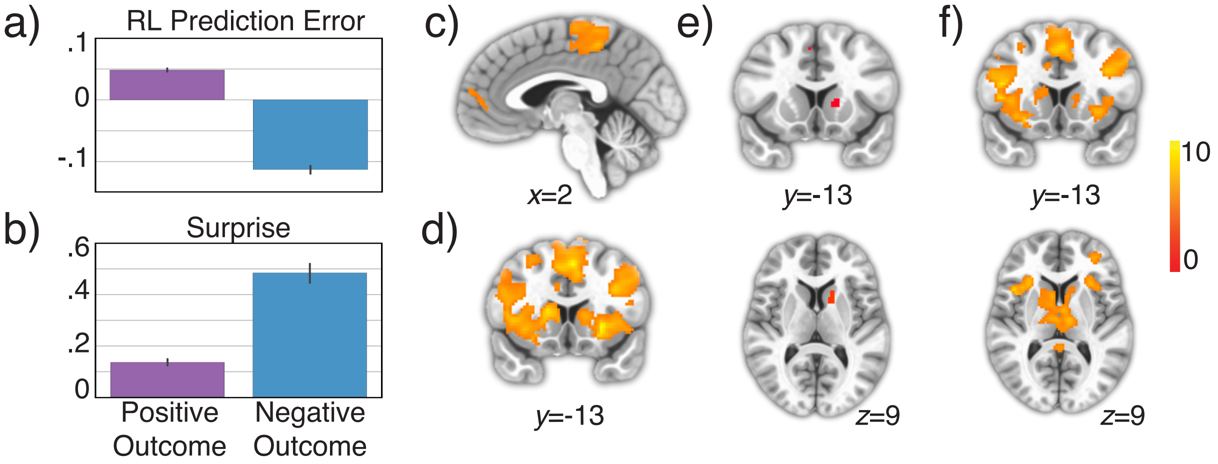

Figure 3: Striatum represents Bayesian Surprise, not reinforcement learning prediction error

a) Mean prediction error from the best fitting RL model, sorted by whether outcome was positive or negative. b) Mean surprise from Bayesian rule learning model, sorted by whether outcome was positive or negative. c) Whole brain corrected results for the contrast of positive > negative outcomes. There were no significant voxels in the striatum for this contrast. d) Whole brain corrected results for the contrast of negative > positive outcomes. e) Results of a conjunction analysis displaying voxels that are significantly active for both negative > positive outcomes and the parametric effect of surprise. Both contrasts were corrected for multiple comparisons across the whole brain before being entered into the conjunction analysis. f) Whole brain corrected results for the contrast of parametric surprise > parametric prediction error, without the effect of outcome partialed out.

1.4 Striatum activation varies with Bayesian surprise

The outcome valence analysis ruled out RL prediction errors as a plausible account of striatal activation in our task. This analysis hinged on the mean effect of valence on error signal. We next turned to a model-based analysis to probe whether trial-by-trial fluctuations in surprise account for striatal activation. We implemented a stringent criterion for detecting surprise in a voxel: it must show a conjunction of a main effect of negative > positive outcomes and a parametric effect of surprise once outcome valence has been partialed out. We used an analogous criterion for RL prediction error. This strict criterion is necessary because outcome valence accounts for a large proportion of the variance attributable to both surprise and prediction error. If this variance isn't accounted for, a parametric surprise regressor could account for significant variability in voxels that are sensitive to some unrelated aspect of outcome, such as the color of the outcome text on the display screen.

We included surprise and prediction error in the same model so that any shared variability would not systematically bias results. Using the conjunction criterion described above, we observed striatal regions sensitive to surprise, in addition to regions in supplementary motor area (SMA; $p <.05$ corrected, Figure 3e). Conversely, we did not observe any significant RL prediction error associated activations at our whole brain threshold, and did not observe any in the striatum even at a lenient threshold ($p<0.01$, uncorrected). Since the conjunction criterion established that the striatal surprise response was not due to outcome alone, we next formally compared surprise and prediction error regressors without projecting out variance due to outcome. This more traditional analysis ruled out any potential impact of removing the outcome variance associated with surprise or prediction error. We observed a robust response in the striatum for the contrast of surprise > prediction error ($p < .05$ corrected, Figure 3f), but did not observe any activation for the reverse contrast in the striatum, even at a lenient threshold ($p<0.01$ uncorrected). We additionally tested surprise > prediction error in the executive striatum ROI and found the same pattern ($t = 4.9$, $p<0.001$, Figure 4b). Together with the results of the outcome analysis, we concluded that in our rule-learning task, striatal activation reflects Bayesian surprise rather than RL prediction error. This result is consistent with our behavioral analysis that established that our task design biased subjects towards a strategy of reasoning over rules rather than the incremental buildup of stimulus-response contingencies.

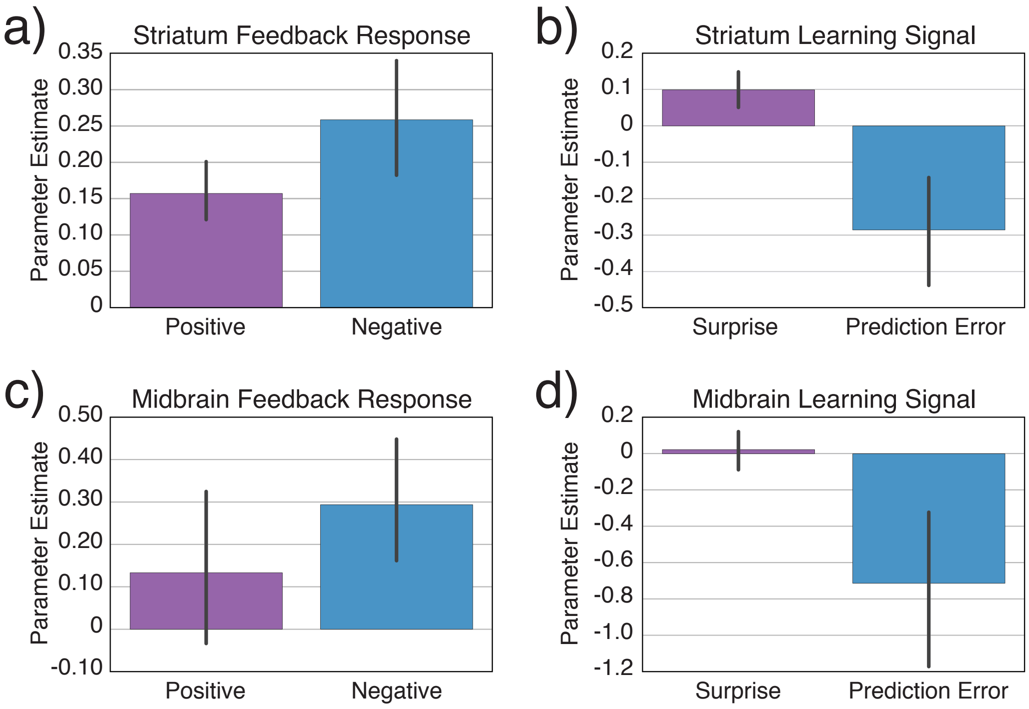

Figure 4: Striatum and Ventral Midbrain ROI Analyses

a) Feedback response to positive and negative outcomes taken from a striatum ROI defined based on its connectivity to executive cortical areas. This pattern is inconsistent with prediction error representation. b) Activation to surprise and RL prediction error taken from the same ROI as in (a). This result indicates that the striatum represents Bayesian surprise rather than prediction error. Note that in this analysis the effect of outcome has not been partialed out, and therefore the negative beta weight for prediction error can arise from the outcome effect in (a). c) Feedback response to positive and negative outcomes taken from a ventral tegmental area ROI. This pattern is inconsistent with prediction error representation. d) Activation to surprise and prediction error taken from the same midbrain ROI. Although the ventral midbrain activation does not support a prediction error signaling account, and its outcome response is consistent with surprise, the parametric effect of surprise was not significant.

1.5 Ventral midbrain activation is inconsistent with prediction error

Midbrain dopamine neurons have been repeatedly shown to signal reward prediction errors, and this property has driven much of the research on the neurobiology of learning. However, midbrain dopamine neurons are heterogeneous, and some respond to surprising outcomes regardless of the valence (Matsumoto and Hikosaka 2009; Bromberg-Martin et al. 2010; Lammel et al. 2012). Such dopamine neurons are hypothesized to be important for a salience signal used for orienting responses. We expected a salience response to roughly correspond to the surprise signal in our rule learning model. We tested wheter ventral midbrain activation was consistent with surprise or prediction error, as evidence for the former would suggest neurons in this region are important for other types of learning.

Ventral midbrain activation was greater for negative compared to positive outcomes (t = 3.66, p=0.003, Figure 4c), indicating that the ventral midbrain signal did not reflect RL prediction errors in our task. Further, surprise provided a better account of ventral midbrain activation than prediction error in a direct comparison (t=3.63, p=0.003). However, in our more stringent test of surprise that separately examined the effect of outcome and the residual variance, residual surprise did not significantly account for ventral midbrain activation (p = 0.56, Figure 4d). We therefore cannot rule out outcome signaling as underlying this effect. Even though fMRI in the ventral midbrain is complicated by many factors (D'Ardenne et al. 2008) and additional brainstem-focused studies are merited, our finding that the ventral midbrain responds more to negative outcomes argues against it being engaged in RL prediction error signaling in this task.

1.6 Rule updating

Our analyses show that the Bayesian model described striatal activation better than RL prediction errors in a rule learning task. However, the striatum also tracks RPE in reinforcement learning tasks(Rutledge et al. 2010). These observations can be reconciled under a model where striatal neurons represent values and these values change in response to errors. In RL tasks, the errors should come from a RPE algorithm, and in a rule-learning task like ours they should reflect rational beliefs about rules. This model suggests that apparent BOLD error signals in the striatum can be better characterized as changes in value signals. Indeed, striatal extracellular dopamine tracks value, rather than prediction error, in a learning task (Hamid et al. 2015). In addition, striatal neurons code stimulus and action values in other domains (Samejima et al. 2005; Lau and Glimcher 2008). However, it is very difficult to distinguish value updating and reward prediction error in existing work that uses RL models of BOLD data because the value update and reward prediction error are perfectly correlated (their ratio is the learning rate). Our Bayesian model allows us to separately examine representation of surprise and rule updating because these quantities are not perfectly correlated in the model (Methods).

We expected that rule updating would involve both the striatum and the caudal inferior frontal sulcus (cIFS), because of the known role of the cIFS in feature-based rule maintenance and execution (Koechlin et al. 2003; Badre and D'Esposito 2007; 2009). Specifically, we hypothesized that if the cIFS maintains and executes the feature-based rule governing behavior, then neural activity in this region should change when the rule governing behavior was more likely to change.

We defined rule updating as the Kullback-Liebler (KL) divergence between the rule probability distributions estimated by the Bayesian model before and after feedback. The KL divergence quantity will be higher when participants were more likely to shift their internal representation of rule likelihoods. Although surprise and rule updating were correlated ($\rho = 0.27$, $p=0.003$), regressors modeling surprise and hypothesis updating responses were entered into the same linear model to ensure that each captured a distinct component of neural activation. We identified brain regions correlated with rule updating at the time of feedback; these regions extended through the bilateral cIFS, intra-parietal lobule (IPL), fusiform gyrus, and portions of the dorsal caudate (Figure 5).

Although rule updating occurs during feedback, it is likely that some updating of rule-response contingencies happens during subsequent cue periods. Evidence for this comes from task-switching paradigms, in which subjects incur a residual switch cost even when they have ample time to prepare for the new task (Sohn et al. 2000; Monsell 2003). We reasoned that rule updating might occur both during feedback and during the cue period of the subsequent trial. We observed activation patterns corresponding to rule updating during the cue period of the subsequent trial in the cIFS and fusiform gyrus bilaterally (Figure 5). These clusters appeared to overlap substantially with the feedback period results; we formally tested this relationship by performing a conjunction analysis to identify the voxels that were active across both maps (Figure 5; (Nichols et al. 2005). Both analyses yielded a statistically significant subregion in the cIFS.

Finally, because the change to rule likelihoods is greater following negative feedback, we wished to exclude outcome valence as a possible mediator of the effects in cIFS. We excluded outcome valence by performing a conjunction test of 1) negative > positive feedback and 2) rule updating, after the effect of valence had been partialed out via hierarchical regression. Again, we observed activation that included the striatum and the cIFS, indicating striatal and cIFS involvement in both the representation of surprise and associated rule updating at the time of feedback. However, only the cIFS (not the striatum) scaled with rule updating during both the feedback and subsequent cue periods. Our findings are consistent with a model in which the cIFS maintains and updates the rules governing behavior, while the striatum maintains and updates the values of the dominant and competing rules.

1.7 Functional connectivity

The results of the preceding analyses suggest that model-based reward learning is facilitated by interactions between the dorsal striatum and cIFS. We expected that functional connectivity between these regions should increase during the feedback period, during which most rule updating was likely to occur. We used the cIFS subregion involved in rule updating during both cue and feedback as the seed region in a psycho-physiological interaction (PPI) analysis. Connectivity between cIFS and dorsal striatum increased during the feedback period (Figure 5A), buttressing our claim that rule updating occurs via interactions between the striatum, which we hypothesize represents the values of rules, and the cIFS, which other literature suggests maintains and executes the most accurate one (Miller and Cohen 2001; Koechlin et al. 2003; Badre and D'Esposito 2009).

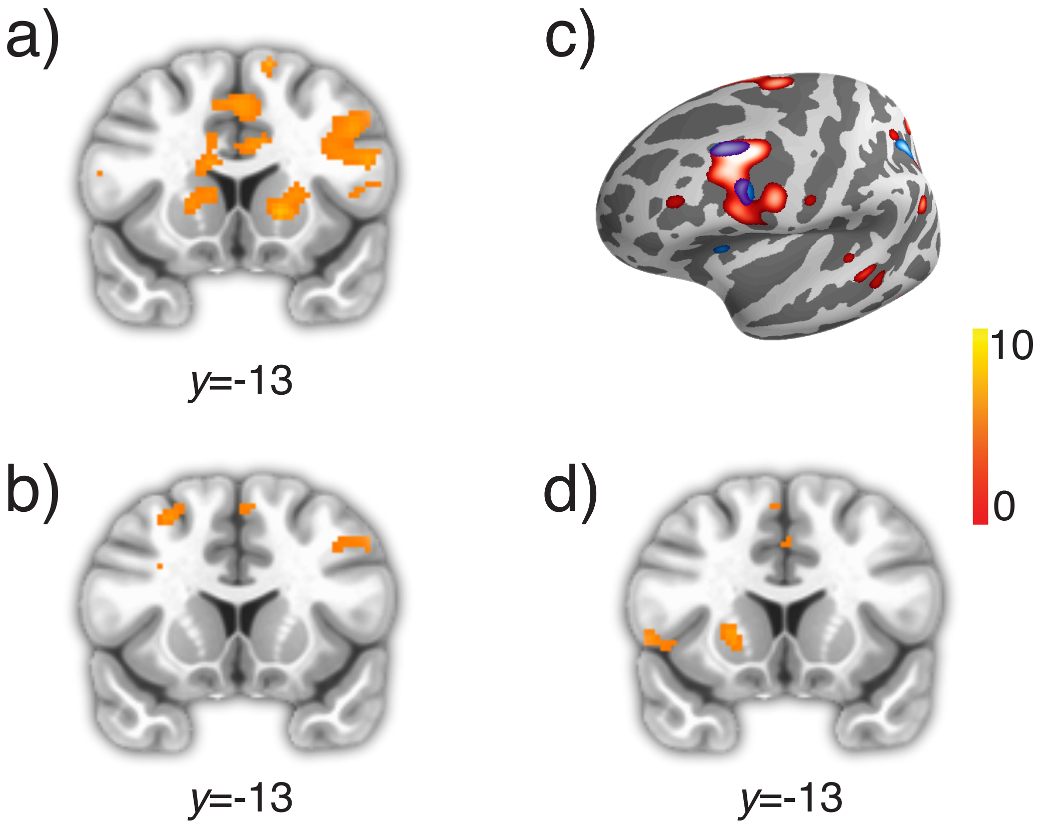

Figure 5: Rule Updating

a) Rule updating during the feedback period in the striatum and left cIFS b) Rule updating during the subsequent cue period in the left cIFS. c) Projections of a and c onto the cortical surface. Red corresponds to rule updating during the feedback period, blue corresponds to rule updating during the following cue period. d) Connectivity analysis showing functional connectivity between the cIFS conjunction cluster and the striatum.

DISCUSSION

The association of the striatal feedback response with reward prediction error saturates the human cognitive neuroscience literature. Several studies have hypothesized that reward prediction error representation is the primary function of striatum during learning (Hare et al. 2008; Daw et al. 2011; Garrison et al. 2013). We found that the striatal feedback response does not reflect prediction error in a rule learning task that does not depend on reinforcement learning. To accommodate both sets of findings, we suggest that the striatal feedback response reflects an update to its representation of values of stimuli, actions or cortical representations. We additionally find that both the striatum and cIFS track the change in beliefs about rules, and that they are functionally coupled during feedback. Together, these results suggest that cortico-striatal interactions support learning about structured relationships.

In order to probe the nature of the striatal update signal outside of the realm of reinforcement learning, we designed a task in which participants are biased towards reasoning about explicit rules, rather than relying on the gradual build-up of stimulus-response contingencies. Our behavioral analysis confirmed that our task design elicited behavior that was consistent with a Bayesian optimal rule learning strategy. We next exploited a difference between the feedback signals generated from RL models (reward prediction error) and Bayesian rule learning (surprise). RL prediction error is a signed learning signal that is larger for positive than negative feedback; in contrast, surprise is larger for negative feedback in our task. Negative feedback generated a stronger response in the striatum, which rules out a RL prediction error account of the striatal feedback response in our task. Further, a striatal BOLD responses tracked a parametric measure of Bayesian surprise. Finally, surprise accounted for striatal responses better than prediction error. Together, these results indicate that the striatal feedback response reflects Bayesian surprise in task conditions where behavior is driven by explicit reasoning about abstract rules.

Striatal feedback responses were not well characterized by RL prediction error in our learning paradigm. Yet, in reinforcement learning tasks, the striatal feedback response clearly tracks reward prediction error (Rutledge et al. 2010). These differences across paradigms can be accounted for if the striatal feedback response reflects the change in values encoded by striatal neurons in response to new information. This more general account of the striatal feedback response is bolstered by several experimental observations. First, striatal response to negative feedback increases if the feedback is less predictable (Lempert and Tricomi 2015). Second, striatal feedback responses are sensitive to subjects' goals, responding differentially to episodic retrieval success and feedback depending on task demands (Han et al. 2010). Third, striatum responds more to negative feedback that indicates a set-shift in the Wisconsin card sorting task (Monchi et al. 2001); however, this observation is less reliable than the typical prefrontal findings

(Buchsbaum et al. 2005; Nyhus and Barceló 2009). Finally, cyclic voltammetry measurements in rodents learning a task indicate that striatal dopamine appears to track a value, rather than prediction error, signal (Hamid et al. 2015). To build upon these experimental observations, we leveraged model-based fMRI and designed a novel paradigm to examine the feedback response under conditions where learning does not depend on the incremental adjustment of stimulus-response contingencies. However, future work measuring the response properties of striatal neurons in humans should directly probe the information carried by these neurons during learning.

We did not use explicit rewards, such as money, in our task, which differs from some RL experiments. However, the striatum is consistently sensitive to feedback in a manner that is similar to its response to explicit rewards (Elliott et al. 1997; Seger and Cincotta 2005; Tricomi et al. 2006; Marco-Pallarés et al. 2007; Dobryakova and Tricomi 2013; Swanson and Tricomi 2014; Lempert and Tricomi 2015). Also, striatum has been shown to respond to internally generated reward prediction errors used in hierarchical RL (Ribas-Fernandes et al. 2011; Diuk et al. 2013; Iglesias et al. 2013). Both empirical and theoretical work suggests that the brain's learning system should use surrogate rewards to learn in the absence of monetary reward receipt.

We not only found that BOLD responses in the striatum were best represented by Bayesian surprise in our task, but we also found that ventral midbrain activation was inconsistent with a prediction error signal. Though surprising, this result agrees with empirical and theoretical work documenting complexity in the dopamine system

(Bromberg-Martin et al., 2010). The prediction error hypothesis of midbrain dopamine function has broad empirical support (Glimcher 2011), including recent optogenetic work (Steinberg et al. 2013; Eshel et al. 2015) but has never been able to account for the full repertoire of midbrain dopamine neuronal firing patterns (Redgrave et al. 1999; Fiorillo et al. 2003; Bromberg-Martin and Hikosaka 2009; Bromberg-Martin et al. 2010; Berridge 2012). This complexity should not be surprising given the profound effects of dopamine on physiology in multiple circuits throughout the brain (Goto et al. 2007). There is widespread empirical data suggesting that prediction error is just one of several signals, including novelty (Lisman and Grace 2005) and motivational salience (Bromberg-Martin et al. 2010), that cause downstream dopamine to modulate circuit properties like signal-to-noise ratio in the prefrontal cortex or plasticity in the hippocampus.

In some studies, striatal BOLD responses reflect prediction errors, but in ours, it reflects Bayesian surprise. We interpret this apparent discrepancy as evidence that striatum is involved in updating internal value representations in response to new information. In RL, prediction error and value updating are perfectly correlated and therefore difficult to distinguish. We leveraged the fact that surprise and rule updating are less correlated in the Bayesian model and were able to show that striatum and cIFS were involved in updating task rules. The cIFS subregion we identified overlaps with the so-called “pre-PMd” (Badre and D'Esposito 2009), which is involved in the maintenance and execution of rules that depend on the features of visual stimuli (Koechlin et al. 2003; Badre and D'Esposito 2007). Our observation that cIFS activation scaled with rule updating agrees with a model in which this area maintains and executes the dominant task rule and that the dominant rule driving behavior is changed based on rule value representations in the striatum. The observation that cIFS and striatum are functionally connected during the feedback period provides support for this model. Future work will investigate the chain of sensory processing that connects visual features to abstract decision rules in cortex, and the hippocampus may play a critical role in this process (Mack et al. 2016).

This study tested the simple idea that the role of the striatum during learning is flexible: the striatum computes the value of potential behavioral policies and updates them in response to new information, including reward prediction errors. We delineated the functional neuroanatomy underlying rule-based learning and in the process ruled out a RL account of striatal activation in deterministic category learning. Our results suggest that value updates in the striatum and cortico-striatal connections facilitate the integration of evidence to guide behavior based on abstract rules.

Acknowledgements

This work was funded by a NeuroVentures pilot grant through Stanford University. The authors would like to thank G. Elliott Wimmer, Vishnu Murty and Kim D'Ardenne for helpful feedback.

REFERENCES

Badre D, D'Esposito M. 2007. Functional Magnetic Resonance Imaging Evidence for a Hierarchical Organization of the Prefrontal Cortex. J Cogn Neurosci. 19:2082–2099.

Badre D, D'Esposito M. 2009. Is the rostro-caudal axis of the frontal lobe hierarchical? Nat Rev Neurosci. 10:659–669.

Ballard IC, Murty VP, Carter RM, MacInnes JJ, Huettel SA, Adcock RA. 2011. Dorsolateral prefrontal cortex drives mesolimbic dopaminergic regions to initiate motivated behavior. J Neurosci. 31:10340–10346.

Berridge KC. 2012. From prediction error to incentive salience: mesolimbic computation of reward motivation. European Journal of Neuroscience. 35:1124–1143.

Bromberg-Martin ES, Hikosaka O. 2009. Midbrain Dopamine Neurons Signal Preference for Advance Information about Upcoming Rewards. Neuron. 63:119–126.

Bromberg-Martin ES, Matsumoto M, Hikosaka O. 2010. Dopamine in Motivational Control: Rewarding, Aversive, and Alerting. Neuron. 68:815–834.

Buchsbaum BR, Greer S, Chang WL. 2005. Meta-analysis of neuroimaging studies of the Wisconsin Card-Sorting task and component processes - Buchsbaum - 2005 - Human Brain Mapping - Wiley Online Library. Human brain ….

Buschman TJ, Denovellis EL, Diogo C, Bullock D, Miller EK. 2012. Synchronous Oscillatory Neural Ensembles for Rules in the Prefrontal Cortex. Neuron. 76:838–846.

D'Ardenne K, McClure SM, Nystrom LE, Cohen JD. 2008. BOLD responses reflecting dopaminergic signals in the human ventral tegmental area. Science. 319:1264–1267.

Daw ND, Gershman SJ, Seymour B, Dayan P, Dolan RJ. 2011. Model-based influences on humans' choices and striatal prediction errors. Neuron. 69:1204–1215.

Daw ND, Niv Y, Dayan P. 2005. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nat Neurosci. 8:1704–1711.

Delgado MR. 2007. Reward-Related Responses in the Human Striatum. Annals of the New York Academy of Sciences. 1104:70–88.

Diuk C, Tsai K, Wallis J, Botvinick M, Niv Y. 2013. Hierarchical learning induces two simultaneous, but separable, prediction errors in human basal ganglia. J Neurosci. 33:5797–5805.

Dobryakova E, Tricomi E. 2013. Basal ganglia engagement during feedback processing after a substantial delay. Cognitive, Affective, & Behavioral Neuroscience. 13:725–736.

Elliott R, Frith CD, Dolan RJ. 1997. Differential neural response to positive and negative feedback in planning and guessing tasks. Neuropsychologia. 35:1395–1404.

Eshel N, Bukwich M, Rao V, Hemmelder V, Tian J, Uchida N. 2015. Arithmetic and local circuitry underlying dopamine prediction errors. Nature. 525:243–246.

Fiorillo CD, Tobler PN, Schultz W. 2003. Discrete Coding of Reward Probability and Uncertainty by Dopamine Neurons. Science. 299:1898–1902.

Garrison J, Erdeniz B, Done J. 2013. Prediction error in reinforcement learning: a meta-analysis of neuroimaging studies. Neuroscience & Biobehavioral Reviews. 37:1297–1310.

Glascher J, Daw N, Dayan P, Doherty JPO. 2010. States versus Rewards: Dissociable Neural Prediction Error Signals Underlying Model-Based and Model-Free

Reinforcement Learning. Neuron. 66:585–595.

Glimcher PW. 2011. Understanding dopamine and reinforcement learning: the dopamine reward prediction error hypothesis. Proc Natl Acad Sci USA. 108 Suppl 3:15647–15654.

Goodman ND, Tenenbaum JB, Feldman J, Griffiths TL. 2008. A rational analysis of rule-based concept learning. Cogn Sci. 32:108–154.

Goto Y, Otani S, GRACE A. 2007. The Yin and Yang of dopamine release: a new perspective. Neuropharmacology. 53:583–587.

Haber SN, Knutson B. 2009. The Reward Circuit: Linking Primate Anatomy and Human Imaging. Neuropsychopharmacology. 35:4–26.

Hamid AA, Pettibone JR, Mabrouk OS, Hetrick VL, Schmidt R, Vander Weele CM, Kennedy RT, Aragona BJ, Berke JD. 2015. Mesolimbic dopamine signals the value of work. Nat Neurosci. 19:117–126.

Han S, Huettel SA, Raposo A, Adcock RA, Dobbins IG. 2010. Functional significance of striatal responses during episodic decisions: recovery or goal attainment? J Neurosci. 30:4767–4775.

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.

Iglesias S, Mathys C, Brodersen KH, Kasper L, Piccirelli M, Ouden den HEM, Stephan KE. 2013. Hierarchical Prediction Errors in Midbrain and Basal Forebrain during Sensory Learning. Neuron. 80:519–530.

Kawagoe R, Takikawa Y, Hikosaka O. 2004. Reward-Predicting Activity of Dopamine and Caudate Neurons—A Possible Mechanism of Motivational Control of Saccadic Eye Movement. J Neurophysiol. 91:1013–1024.

Koechlin E, Ody C, Kouneiher F. 2003. The Architecture of Cognitive Control in the Human Prefrontal Cortex. Science. 302:1181–1185.

Lammel S, Lim BK, Ran C, Huang KW, Betley MJ, Tye KM, Deisseroth K, Malenka RC. 2012. Input-specific control of reward and aversion in the ventral tegmental area. Nature. 491:212–217.

Lau B, Glimcher PW. 2008. Value Representations in the Primate Striatum during Matching Behavior. Neuron. 58:451–463.

Lempert KM, Tricomi E. 2015. The Value of Being Wrong: Intermittent Feedback Delivery Alters the Striatal Response to Negative Feedback. J Cogn Neurosci. 28:261–274.

Lisman JE, Grace AA. 2005. The Hippocampal-VTA Loop: Controlling the Entry of Information into Long-Term Memory. Neuron. 46:703–713.

Mack ML, Love BC, Preston AR. 2016. Dynamic updating of hippocampal object representations reflects new conceptual knowledge. Proceedings of the National Academy of Sciences. 113:13203–13208.

Marco-Pallarés J, Müller SV, Münte TF. 2007. Learning by doing: an fMRI study of feedback-related brain activations. NeuroReport. 18:1423–1426.

Matsumoto M, Hikosaka O. 2009. Two types of dopamine neuron distinctly convey positive and negative motivational signals. Nature. 459:837–841.

Miller EK, Cohen JD. 2001. An Integrative Theory of Prefrontal Cortex Function. Annu Rev Neurosci. 24:167–202.

Monchi O, Petrides M, Petre V, Worsley K, Dagher A. 2001. Wisconsin Card Sorting revisited: distinct neural circuits participating in different stages of the task identified by event-related functional magnetic resonance imaging. J Neurosci. 21:7733–7741.

Monsell S. 2003. Task switching. Trends in Cognitive Sciences. 7:134–140.

Montague PR, Dayan P, Sejnowski TJ. 1996. A framework for mesencephalic dopamine systems based on predictive Hebbian learning. J Neurosci. 16:1936–1947.

Morris G, Schmidt R, Bergman H. 2010. Striatal action-learning based on dopamine concentration. Exp Brain Res. 200:307–317.

Murty VP, Shermohammed M, Smith DV, Carter RM, Huettel SA, Adcock RA. 2014. Resting state networks distinguish human ventral tegmental area from substantia nigra. Neuroimage. 100:580–589.

Nichols T, Brett M, Andersson J, Wager T, Poline J-B. 2005. Valid conjunction inference with the minimum statistic. Neuroimage. 25:653–660.

Niv Y. 2009. Reinforcement learning in the brain. Journal of Mathematical Psychology. 53:139–154.

Niv Y, Daniel R, Geana A, Gershman SJ, Leong YC, Radulescu A, Wilson RC. 2015. Reinforcement Learning in Multidimensional Environments Relies on Attention Mechanisms. J Neurosci. 35:8145–8157.

Nyhus E, Barceló F. 2009. The Wisconsin Card Sorting Test and the cognitive assessment of prefrontal executive functions: A critical update. Brain and Cognition. 71:437–451.

O'Doherty JP, Dayan P, Friston K, Critchley H, Dolan RJ. 2003. Temporal difference models and reward-related learning in the human brain. Neuron. 38:329–337.

O'Reilly JX, Schüffelgen U, Cuell SF, Behrens TE, Mars RB, Rushworth MF. 2013. Dissociable effects of surprise and model update in parietal and anterior cingulate cortex. Proceedings of the National Academy of Sciences. 110:E3660–E3669.

O’Reilly RC, Rudy JW. 2001. Conjunctive representations in learning and memory: principles of cortical and hippocampal function. Psychol Rev. 108:311.

Piantadosi ST. 2011. Learning and the language of thought. Massachusetts Institute of Technology.

Piantadosi ST, Tenenbaum JB, Goodman ND. 2012. Bootstrapping in a language of thought: A formal model of numerical concept learning. Cognition. 123:199–217.

Redgrave P, Prescott TJ, Gurney K. 1999. Is the short-latency dopamine response too short to signal reward error? Trends in Neurosciences. 22:146–151.

Reynolds JN, Hyland BI, Wickens JR. 2001. A cellular mechanism of reward-related learning. Nature. 413:67–70.

Ribas-Fernandes JJF, Solway A, Diuk C, McGuire JT, Barto AG, Niv Y, Botvinick MM. 2011. A Neural Signature of Hierarchical Reinforcement Learning. Neuron. 71:370–379.

Rigoux L, Stephan KE, Friston KJ, Daunizeau J. 2014. Bayesian model selection for group studies - revisited. Neuroimage. 84:971–985.

Rutledge RB, Dean M, Caplin A, Glimcher PW. 2010. Testing the Reward Prediction Error Hypothesis with an Axiomatic Model. J Neurosci. 30:13525–13536.

Samejima K, Ueda Y, Doya K, Kimura M. 2005. Representation of action-specific reward values in the striatum. Science. 310:1337–1340.

Schultz W. 1997. A Neural Substrate of Prediction and Reward. Science. 275:1593–1599.

Seger CA, Cincotta CM. 2005. The roles of the caudate nucleus in human classification learning. J Neurosci. 25:2941–2951.

Sohn MH, Ursu S, Anderson JR, Stenger VA, Carter CS. 2000. The role of prefrontal cortex and posterior parietal cortex in task switching. Proceedings of the National Academy of Sciences. 97:13448–13453.

Steinberg EE, Keiflin R, Boivin JR, Witten IB, Deisseroth K, Janak PH. 2013. A causal link between prediction errors, dopamine neurons and learning. Nat Neurosci. 16:966–973.

Stephan KE, Penny WD, Daunizeau J, Moran RJ, Friston KJ. 2009. Bayesian model selection for group studies. Neuroimage. 46:1004–1017.

Sutton RS, Barto AG. 1998. Introduction to Reinforcement Learning. MIT Press.

Swanson SD, Tricomi E. 2014. Goals and task difficulty expectations modulate striatal responses to feedback. Cognitive, Affective, & Behavioral Neuroscience. 14:610–620.

Tenenbaum JB, Kemp C, Griffiths TL, Goodman ND. 2011. How to Grow a Mind: Statistics, Structure, and Abstraction. Science. 331:1279–1285.

Tricomi E, Delgado MR, McCandliss BD, McClelland JL, Fiez JA. 2006. Performance Feedback Drives Caudate Activation in a Phonological Learning Task. http://dxdoiorgezproxystanfordedu/101162/jocn20061861029. 18:1029–1043.

Tziortzi AC, Haber SN, Searle GE, Tsoumpas C, Long CJ, Shotbolt P, Douaud G, Jbabdi S, Behrens TEJ, Rabiner EA, Jenkinson M, Gunn RN. 2013. Connectivity-Based Functional Analysis of Dopamine Release in the Striatum Using Diffusion-Weighted MRI and Positron Emission Tomography. Cereb Cortex. 24:bhs397–bhs1177.