Behavioral Neuroscience

Reinforcement Learning Modeling Reveals a Reward-History-Dependent Strategy Underlying Reversal Learning in Squirrel Monkeys

Bilal A. Bari, Megan J. Moerke, Hank P. Jedema, Devin P. Effinger, Jeremiah Y. Cohen, and Charles W. Bradberry

Online First Publication, September 27, 2021. http://dx.doi.org/10.1037/bne0000492

CITATION

Bari, B. A., Moerke, M. J., Jedema, H. P., Effinger, D. P., Cohen, J. Y., & Bradberry, C. W. (2021, September 27). Reinforcement Learning Modeling Reveals a Reward-History-Dependent Strategy Underlying Reversal Learning in Squirrel Monkeys. *Behavioral Neuroscience*. Advance online publication. http://dx.doi.org/10.1037/bne0000492

Behavioral Neuroscience

© 2021 American Psychological Association
ISSN: 0735-7044

https://doi.org/10.1037/bne0000492

Reinforcement Learning Modeling Reveals a Reward-History-Dependent Strategy Underlying Reversal Learning in Squirrel Monkeys

Bilal A. Bari[1, 2], Megan J. Moerke[3], Hank P. Jedema[3], Devin P. Effinger[4], Jeremiah Y. Cohen[1, 2], and Charles W. Bradberry[3]

[1] The Solomon H. Snyder Department of Neuroscience, Brain Science Institute, Johns Hopkins University
[2] Kavli Neuroscience Discovery Institute, Johns Hopkins University
[3] NIDA Intramural Research Program, Baltimore, Maryland, United States
[4] Department of Pharmacology, University of North Carolina at Chapel Hill

Insight into psychiatric disease and development of therapeutics relies on behavioral tasks that study similar cognitive constructs in multiple species. The reversal learning task is one popular paradigm that probes flexible behavior, aberrations of which are thought to be important in a number of disease states. Despite widespread use, there is a need for a high-throughput primate model that can bridge the genetic, anatomic, and behavioral gap between rodents and humans. Here, we trained squirrel monkeys, a promising preclinical model, on an image-guided deterministic reversal learning task. We found that squirrel monkeys exhibited two key hallmarks of behavior found in other species: integration of reward history over many trials and a side-specific bias. We adapted a reinforcement learning model and demonstrated that it could simulate squirrel monkey-like behavior, capture training-related trajectories, and provide insight into the strategies animals employed. These results validate squirrel monkeys as a model in which to study behavioral flexibility.

Keywords: reversal learning, reinforcement learning, squirrel monkeys, decision making, behavioral modeling

Supplemental materials: https://doi.org/10.1037/bne0000492.supp

Psychiatry is in need of fundamental insights both so we may understand the psychological and neural basis of disease and so we can develop novel therapeutics. Comparative neuroscience is one approach that has historically led to the discovery of safe and effective pharmaceuticals (Markou et al., 2009). One common procedure is to combine animal models of psychiatric disease with commonly used behavioral tasks, such as the forced swim test or elevated plus maze. Compounds can then be tested for their ability to ameliorate modeled symptoms (Crawley, 2008; Dawson & Tricklebank, 1995; Flint & Shifman, 2008; Kaiser & Feng, 2015; Nestler et al., 2002). While this approach has been fruitful, it has failed to fundamentally change our understanding about disease or lead to the discovery of truly novel therapeutics (Fenton et al., 2003; Pangalos et al., 2007). This is partly because preclinical behavioral models represent a major bottleneck in drug development (Tallman, 1999). Since psychiatric diseases affect higher order cognitive processes, designing tasks that translate drug effects from animal models to humans is nontrivial. The standard library of tasks was designed to probe intuitive ideas about observable symptoms, not quantitative theories of cognitive processes.

A promising approach is to design behavioral tasks that probe the same psychological phenomena across species (Pike et al., 2021). These tasks may use different stimuli and motor responses, but attempt to isolate the same neurocomputational mechanisms.

Bilal A. Bari https://orcid.org/0000-0002-8381-3802
Hank P. Jedema https://orcid.org/0000-0002-7869-2110

We thank Samantha O. Brown and Rodrigo A. Montoro for assistance with monkey training. We thank Frederick Barrett, Matthew Johnson, Sandeep Nayak, Manoj Doss, Justin Strickland, Ceyda Sayalı, and Cooper Grossman for valuable discussions at various stages of the project. We thank Cooper Grossman for providing feedback on an earlier version of the manuscript.

This work was supported by R01DA042038 (Jeremiah Y. Cohen and Bilal A. Bari), R01NS104834 (Jeremiah Y. Cohen and Bilal A. Bari), T32GM007309 (Bilal A. Bari), and the NIDA Intramural Research Program (Megan J. Moerke, Hank P. Jedema, Devin P. Effinger, and Charles W. Bradberry).

We have no conflicts of interest to disclose.

Bilal A. Bari played lead role in formal analysis, methodology, software, visualization and writing of original draft and equal role in conceptualization, and writing of review and editing. Megan J. Moerke played supporting role in data curation, investigation, methodology, and writing of review and editing. Hank P. Jedema played supporting role in data curation, investigation, methodology, and writing of review and editing. Devin P. Effinger played supporting role in data curation, investigation, and methodology. Jeremiah Y. Cohen played supporting role in conceptualization, supervision, and writing of review and editing. Charles Bradberry played lead role in funding acquisition, project administration, resources and supervision and supporting role in conceptualization, and writing of review and editing.

Correspondence concerning this article should be addressed to Charles W. Bradberry, NIDA Intramural Research Program, 251 Bayview Boulevard, Suite 200, Baltimore, MD 21224, United States. Email: charles.bradberry@nih.gov

Theories about the relevant cognitive processes are used to form and test explicit hypotheses, in the form of models, about behavioral strategies (Daw, 2011; Heathcote et al., 2015; Wilson & Collins, 2019). From the perspective of computational psychiatry, these models in turn allow us to understand and quantify aberrant information processing in disease (Aylward et al., 2019; Gershman & Lai, 2021; Huys et al., 2016; Mason et al., 2017; Radulescu & Niv, 2019; Redish, 2004), as well as effects of therapy (Frank et al., 2007; Michely et al., 2020; Paulus et al., 2016).

Reversal learning is one popular task that is amenable to theory-based computational modeling (Behrens et al., 2007; Soltani & Izquierdo, 2019). In common variants of this task, subjects are presented with a choice between two stimuli, one associated with a high-value outcome (e.g., high probability or large volume of reward) and the other associated with a low-value outcome. Subjects begin the task with no knowledge about which stimulus is the better option. On each trial, subjects select one stimulus, receive the associated outcome, and repeat this process. Through trial and error, subjects learn the values of each stimulus. After some time, the two stimuli reverse in association (hence, reversal learning), so that the previously low-value stimulus becomes the high-value stimulus and vice versa. Importantly, these reversals are not cued, necessitating continual trial-by-trial learning to maximize reward. This task design is thought to engage mechanisms of flexible and rapid learning, impairments of which are implicated in a wide range of psychiatric disease (Aylward et al., 2019; Brigman et al., 2009; Huys et al., 2013; Izquierdo & Jentsch, 2012; Leeson et al., 2009; Remijnse et al., 2006; Swainson et al., 2000), including addiction (Ersche et al., 2011; Porter et al., 2011).

Although behavior on these tasks is typically reported using simple summary statistics (average performance, trials to reach a criterion, etc.), richer insight can be gleaned with reinforcement learning modeling. Reinforcement learning is a framework that formalizes learning from environmental feedback (Sutton & Barto, 1998) and has provided a number of tractable algorithms that have delineated numerous structure–function relationships in the nervous system (Bari et al., 2019; Grossman et al., 2020; O'Doherty et al., 2004; Ottenheimer et al., 2020; Samejima et al., 2005; Schultz et al., 1997). Among the most commonly applied algorithms are the class that iteratively learn stimulus values over many trials and choose based on the relative values of the stimuli. Reinforcement learning models of reversal learning have been used to explain behavioral data in species as diverse as rodents (Harris et al., 2020; Metha et al., 2020), macaques (Costa et al., 2016), and humans (Kanen et al., 2019). Importantly, reinforcement learning models are generative models—that is, they are capable of simulating behavior, a premise which we capitalize on in this manuscript.

Here, we trained squirrel monkeys on a deterministic image-based reversal learning task. Squirrel monkeys are New World primates widely used in biomedical research, primarily due to their small size (<1 kg), ease of handling, and adaptation to laboratory conditions (Abee, 2000). From the perspective of comparative neuroscience, squirrel monkeys help span the massive genetic, anatomical, and behavioral gap between rodents and humans (Boinski, 1999). They may also prove to be a useful preclinical model for development of optogenetic-based interventions (O'Shea et al., 2018).

Our objective was to determine if squirrel monkeys solve reversal learning tasks using a strategy compatible with trial-by-trial reinforcement learning and to isolate parameters of cognitive flexibility to employ in future studies. First, we demonstrate that squirrel monkeys do not adopt the optimal win–stay/lose–shift strategy required to optimize reward accumulation in this task but rather integrate reward over many trials. We fit a number of reinforcement learning models and found that a standard Rescorla–Wagner model fit best, similar to reversal learning models in other species, including rhesus macaques (Costa et al., 2016). We show that this model simulates realistic behavior, providing a convincing platform for making inferences about behavioral strategy. Finally, we use the recovered parameters to define how the behavioral strategy develops with training. We primarily analyze data at the monkey level to allow one to trace an individual monkey throughout the manuscript and observe how well the reinforcement learning models capture monkey-to-monkey variability, since accounting for this variability is critical for future studies.

Method

Subjects

A total of 13 (9 of which met behavioral criteria) adult male squirrel monkeys (*Saimiri sciureus*) with less than 1 year of training on behavioral touchscreen tasks were housed individually under controlled temperature and humidity on a 12/12-hr light–dark cycle (lights on from 0700 to 1900 hr). Monkeys weighed 867–1,113 g (*M*: 965 g) and were maintained on a diet of primate chow (LabDiet High Protein Monkey Biscuits; PMI Feeds, St. Louis, MO) with continuous access to water in the home chamber. Environmental enrichment, including fresh fruits and vegetables, was provided on a daily basis. The maintenance and experimental use of animals was carried out in accordance with the 2011 Guide for Care and Use of Laboratory Animals. All experimental protocols were approved by the Animal Care and Use Committee of the National Institute on Drug Abuse Intramural Research Program.

Apparatus

Experiments were conducted in sound-attenuating chambers equipped with a 15″ touchscreen (Elo TouchSystems, Menlo Park, CA), mounted in a panel 14.25″ from the floor of the chamber. Centered 1.5″ below the touchscreen and extending 2″ into the chamber was a well into which measured volumes of 30% sweetened condensed milk (Eagle Foods, Richfield, OH) could be delivered through a line connected to a syringe pump (Harvard Apparatus, South Natick, MA) located outside the chamber. Monkeys were seated in custom-built acrylic chairs facing the touchscreen panel. A computer and software program (E-Prime Professional 3.0; Psychology Software Tools, Inc., Sharpsburg, PA) controlled the parameters of the experimental program and data collection.

Behavioral Task and Training

Naïve squirrel monkeys were single housed and handled using chain and collar methods (Kelleher et al., 1963). Animals were acclimated to custom designed and fabricated behavioral chairs in sound-attenuating chambers in which they were restrained only at the waist, leaving their upper bodies unrestrained. Monkeys were first trained to touch a yellow square stimulus presented at the center of the touchscreen for delivery of a 30% sweetened condensed milk dilution. A correct response resulted in delivery of a fixed volume of 0.15 mL/kg of milk and initiated an intertrial interval (ITI) of 2 s during which responses on the screen had no programmed consequence. As animals learned to respond precisely, the ITI was increased and the stimulus was gradually reduced in size and presented at random locations on the screen. The time required for this familiarization training was 12 weeks + 0.6 (standard error of the mean; SEM). Prior to introducing the reversal task, monkeys were then trained on a task in which they chose between different quantities of milk (0.075–0.3 mL/kg) represented by unique stimuli on the touchscreen. Monkeys registered a choice by physically touching the display for 100 ms; this was paired with a tone and allocation of the associated reward into the well below the touchscreen. A house light and speakers inside the experimental chambers provided illumination and white noise. Training sessions were generally carried out 5 days a week (Monday–Friday) and lasted 30–60 min. Animals performed this task for 18 + 0.2 weeks. This duration was not required for learning the task, but because it provided continued handling during a delay necessitated by events external to this study.

Following the above training and choice tasks, monkeys conducted an image-based deterministic reversal learning task. These sessions began with a Discrimination block, in which monkeys were presented with two novel images selected randomly from a large library of images. One image was associated with a big reward (large volume of milk, the "correct" choice) and the other image was associated with a small reward (small volume of milk, the "incorrect" choice). Monkeys registered a choice by physically touching one of the visual stimuli on the display and received the associated reward from a reward port. Following an 8–12 s ITI, the images were presented again on the next trial, with left/right positions randomized between trials. Monkeys made choices until they reached a performance threshold of 80% correct in the past 15 trials, after a minimum 20 trial block length. Once this threshold was reached, a Reversal block was initiated, in which the two image associations reversed so the image previously associated with big reward was now associated with small reward, and vice versa for the other image. Monkeys again performed until they reached the performance threshold, at which point a new Discrimination block was initiated and two new images were randomly sampled from the library. Monkeys typically performed for 150 trials, although some sessions were shorter due to reduced motivation. The large reward (0.13–0.24 mL/kg) was four times larger than the small reward (0.03–0.06 mL/kg).

Data Analysis

All 13 monkeys completed at least 60 sessions and at least one block per session on average. Monkeys that reached an average performance threshold of 54% across all reward blocks and all sessions were included, yielding nine monkeys in the final data set. Monkeys performed an average of 121 sessions (range 66–135).

All choices that yielded big (small) reward were labeled as correct (incorrect). Performance was defined as the fraction of correct choices in a session. To generate reward history regressions, we arbitrarily coded one image as "Image 0" and the other image as "Image 1" for each set of presented images and fit the following random effects logistic regression

$$\log\left(\frac{P(c_1(t))}{1-P(c_1(t))}\right) = \sum_{i=1}^{15} \beta_i^r (R_1(t-i) - R_0(t-i)) + \sum_{i=1}^{15} \beta_i^S (S(t-i)) + \beta_{\text{int}},$$

where $c_1(t) = 1$ for a choice to "Image 1" and 0 for a choice to "Image 0." $R(t) = 1$ if big reward was delivered for that image on trial t and 0 otherwise. $S(t) = 1$ if the animal chose the rightward side on that trial and 0 if the leftward side was chosen. We included monkey-level and session-level (nested within monkey) random effects for the intercept.

We generated error bars for performance within blocks (Figures 2 D and 4D) by computing bootstrapped 95% confidence intervals from 1,000 bootstrap samples of the mean.

In generating image-based win–stay/lose–shift and mutual information metrics, we excluded the first trial of each Discrimination block, as new images were presented on these trials. The mutual information between stay/shift (to image) and reward on the previous trial was calculated as

$$\begin{aligned}
I(R,S) &= H(S) - H(S|R) \\
H(S) &= -\sum_{s\in S} P(s)\log_2(P(s)) \\
&= -(P(\text{switch})\log_2(P(\text{switch})) + P(\text{stay})\log_2(P(\text{stay}))) \\
H(S|R) &= \sum_{r\in R} H(S|r)P(r) \\
&= -(P(\text{switch}|\text{win})\log_2(P(\text{switch}|\text{win})) \\
&\quad + P(\text{stay}|\text{win})\log_2(P(\text{stay}|\text{win})))P(\text{win}) \\
&\quad - (P(\text{switch}|\text{loss})\log_2(P(\text{switch}|\text{loss})) \\
&\quad + P(\text{stay}|\text{lose})\log_2(P(\text{stay}|\text{lose})))P(\text{lose}),
\end{aligned}$$

where $I(R, S)$ is the mutual information, $S = \{\text{switch, stay}\}$ on the current trial, and $R = \{\text{win, lose}\}$ on the previous trial.

Side bias was defined as $2 \cdot |\frac{N_r}{N_r+N_l} - 0.5|$, where N_r and N_l are the total rightward and leftward choices in a session, respectively. Side bias = 0 if there are an equal number of leftward/rightward choices and 1 if all choices are exclusively to one side.

All regressions relating behavioral metrics to sessions number were random effects linear regressions with monkey-level random effects for slope and intercept.

Reinforcement Learning Models

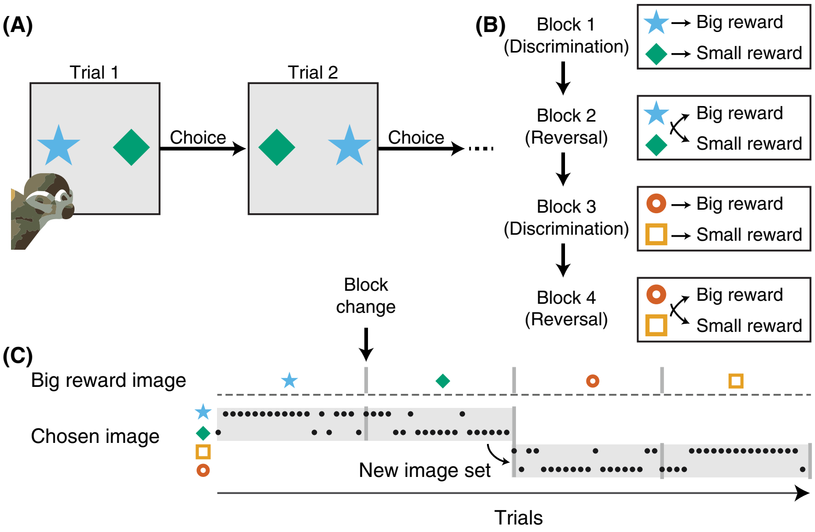
We developed a number of reinforcement learning models based on the Rescora–Wagner model. Our chosen model took the following form for updating image values

$$\begin{aligned}
\delta(t) &= R(t) - V_{\text{chosen}}(t) \\
V_{\text{chosen}}(t+1) &= V_{\text{chosen}}(t) + \alpha \cdot \delta(t) \\
V_{\text{unchosen}}(t+1) &= V_{\text{unchosen}}(t),
\end{aligned}$$

where values were initialized with $V_i = 0$ at the beginning of each Discrimination block. In this model, the chosen image's value is

Figure 1

Reversal Learning Task Design



Note. (A) Squirrel monkeys chose between two images presented on the left and right halves of a touchscreen. A choice was registered by physically touching either visual stimulus on the display. One image was deterministically associated with big milk reward, and the other image was associated with small milk reward. Image locations were randomly displayed on the left and right halves of the screen on separate trials. (B) Monkeys performed sequences of Discrimination and Reversal blocks. At the beginning of each Discrimination block, two new images were randomly sampled from a large library of images and each image was randomly assigned to big or small reward. At the beginning of each Reversal block, the two images switched reward contingencies. Block transitions were triggered by a threshold of 80% correct responses (response to the big reward image) in the past 15 trials, after a minimum of 20 trials. These transitions were unsignaled, requiring the animal to use reward feedback to guide decisions. (C) Example choice behavior demonstrates the flexibility of behavior at Discrimination → Reversal and Reversal → Discrimination block transitions.

updated based on the discrepancy between prediction and reward (reward prediction error, δ(t)). The unchosen image's value remains unchanged. Image values were fed into a softmax function to generate choices according to

$$P(c(t) = \text{rightward}) = \frac{1}{1 + e^{-\beta(V_{\text{rightward}} - V_{\text{leftward}}) - \text{bias}}}$$

$$P(c(t) = \text{leftward}) = 1 - P(c(t) = \text{rightward}).$$

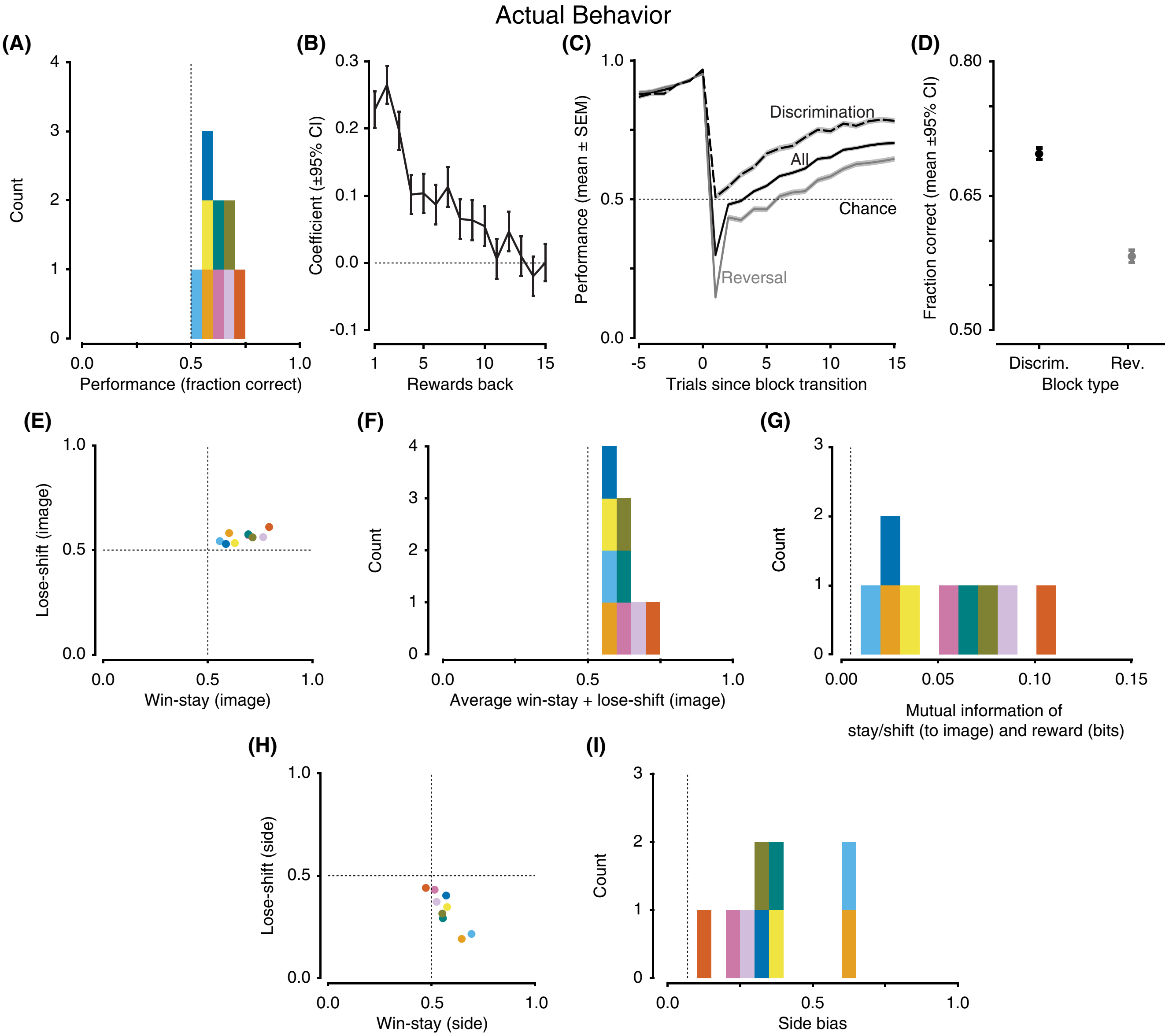
We fit a number of model variants. First, we considered a set of noise models testing whether behavior could be explained as random, biased, perseverative (1-trial-back choice autocorrelation), or biased + perseverative. Within the space of Rescorla–Wagner models, we considered models with all possible combinations of the following: one learning rate, two learning rates (separate learning of positive and negative reward prediction errors), forgetting of unchosen image values to 0, rewards coded as [0.25 1] (since the small reward was 25% the volume of the large reward; only for models with forgetting), and learning of action values (i.e., learning values for leftward/rightward actions). Each of these models also included permutations for nuisance parameters (none, side bias, perseveration, or side bias + perseveration). We additionally considered a set of models that augmented each model to allow for a mixture of image-based win–stay/lose–shift and reinforcement learning. We considered one final model (a variant of our chosen model) that explicitly accounted for a reversal mechanism. In this model, reward prediction errors symmetrically updated image values—if one image value was increased by δ(t), the other image value was decreased by δ(t), while bounding image values by [0 1]. In all models, reward values were coded as 0 and 1 for small reward and big reward, respectively (except for the variant where they were coded as [0.25 1]). In total, we considered 105 model variants. Variations of these models and justifications for these parameters have been reported previously (Bari et al., 2019; Dorfman et al., 2019; Grossman et al., 2020; Kanen et al., 2019; Katahira, 2015).

We developed a metric, the maximal trial-by-trial change in P(choice), to capture the interaction between the learning rate, α, and the inverse temperature, β (Figure 6F, J and Figure S9C). For a given α and β, we assumed the largest reward prediction error, δ(t) = 1. This yields $V_{\text{chosen}}(t + 1) = V_{\text{chosen}}(t) + \alpha \cdot \delta(t)$, which can be simplified as $V_{\text{chosen}}(t + 1) - V_{\text{chosen}}(t) = \alpha$. In other words, the value function is increased by α in response to the largest possible reward prediction error in this task. We then calculated the change in P(choice) around the inflection point of the softmax function [at P(choice) = 0.5] since the slope is steepest at this point. This yields the maximum trial-by-trial change in P(choice).

$$\Delta P(\text{choice}) = \frac{1}{1 + e^{-\beta \cdot \alpha/2}} - \frac{1}{1 + e^{-\beta \cdot -\alpha/2}}.$$

Figure 2

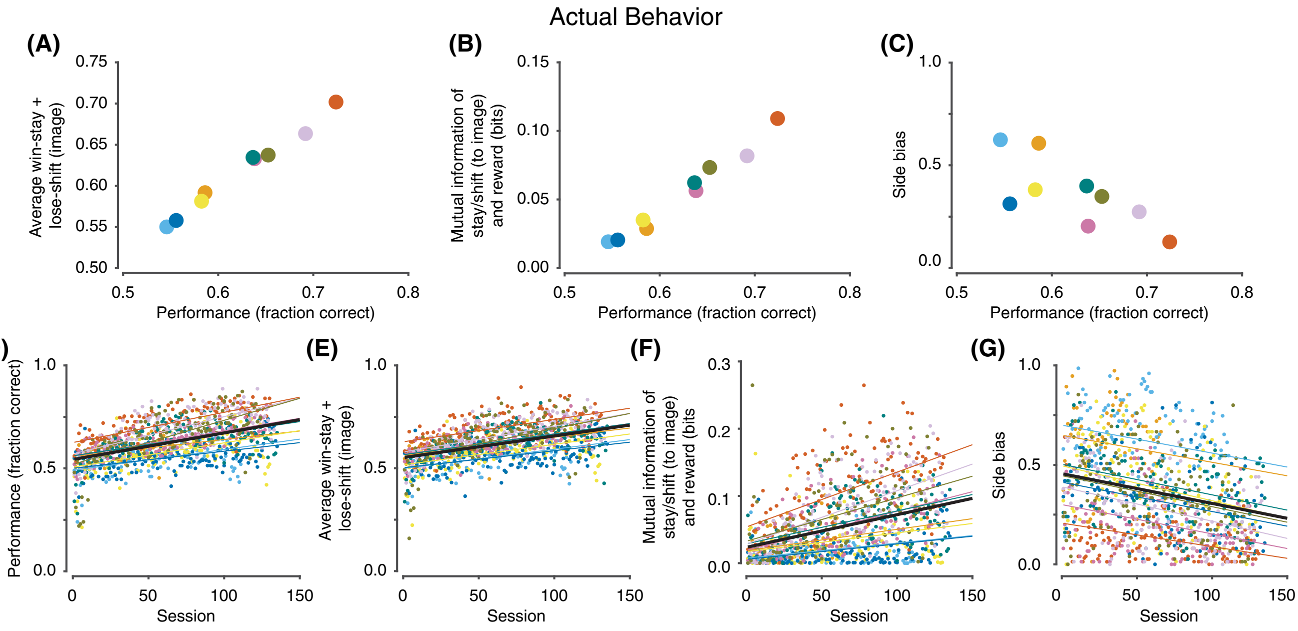
Behavioral Features Demonstrate Reward Sensitivity and Side Bias



Note. (A) Performance was significantly better than chance (50%, dashed line). (B) Logistic regression coefficients for choice as a function of reward history and side history. (C) Performance at block transitions for all blocks, and separately for Discrimination and Reversal blocks. Relative to Reversal blocks, monkeys were faster to improve performance during new Discrimination blocks. The increase in performance prior to block transitions is because transitions were triggered by good performance. (D) Performance was better in Discrimination blocks relative to Reversal blocks. (E) Image-based win–stay and lose–shift were both greater than 0.5, demonstrating that animals learned from both wins (big reward) and losses (small reward) to guide decisions. (F) The average win–stay + lose–shift, which can be taken as a proxy for the strength of reward-guided behavior, was greater than 0.5 (dashed line). Values close to 0.5 are consistent with reward-insensitive behavior and values of 1.0 are consistent with a perfect win–stay lose–shift strategy. (G) The mutual information between stay/switch and reward on the previous trial. Mutual information quantifies how much better we can predict the strategy (stay vs. switch) if we know the reward received on the previous trial (dashed line is from simulated random behavior). (H) Side-based win–stay and lose–shift highlight a side bias, where animals largely stay. (I) Side bias, which is 1 if choices are exclusively to one side and 0 if they are uniformly split, was widely distributed (dashed line is from simulated non-side-biased behavior). Colors denote individual monkeys and are consistent between figures. In panels A, E–I, each data point is the average for one monkey, across all sessions and blocks. Panels B–D are analysis of all the data, pooled across all monkeys, sessions, and blocks.

Figure 3

Relationship Between Performance, Reward Sensitivity, Side Bias, and Training



Note. (A) The average win–stay + lose–shift increased with increased performance. (B) The mutual information between stay/shift and reward increased with performance >0.5. (C) Side bias was higher when performance was closer to 0.5 and reduced when performance was better. (D) Performance improved with more sessions performed. (E) The average win–stay + lose–shift improved with training. (F) The mutual information between stay/switch and reward increased with training. (G) Side bias decreased with training. Black line shows the fixed effect and thin colored lines show individual monkey random effects. Colors denote individual monkeys and are consistent between figures. In panels A–C, each dot is the average for one monkey, across all sessions and blocks. In panels D–G, each dot is the average for one session, across all blocks.

Model Fitting

Models were fit to individual session data with maximum likelihood estimation, with 10 starting points to avoid finding local minima. To determine which models fit the data most parsimoniously, we used the Bayesian information criterion, which penalizes models with additional parameters. The above reinforcement learning model fit best for the most monkeys (Table S1).

This model is notable for several reasons. First, none of the noise models fit best for the monkeys included in this data set. Because Bayesian information criterion, relative to other metrics (like Akaike information criterion), favors simpler models, this suggests that Rescorla–Wagner-like learning is a key feature of behavior. Second, although reinforcement learning models with two learning rates are often fit to animal and human data, we found that none of the models with two learning rates were selected. Third, none of the win–stay/lose–shift models provided better fits than the complementary model variants without win–stay/lose–shift. This includes four win–stay/lose–shift models that did not include Rescorla–Wagner-style learning (a mixture of the four noise models + win–stay/lose–shift).

Model Recovery

For the best model, we took the parameter estimates for each session and generated fictive data according to the same model. We then fit all 105 models to this synthetic data set and found that the true generative model was selected for nine of the nine simulated monkeys. This shows that our model recovery procedure could indeed recover our chosen model.

We also conducted a parameter recovery exercise with these models and found that the difference between actual and recovered parameters had a mode of 0 for all three parameters.

Hierarchical Bayesian Model Fitting

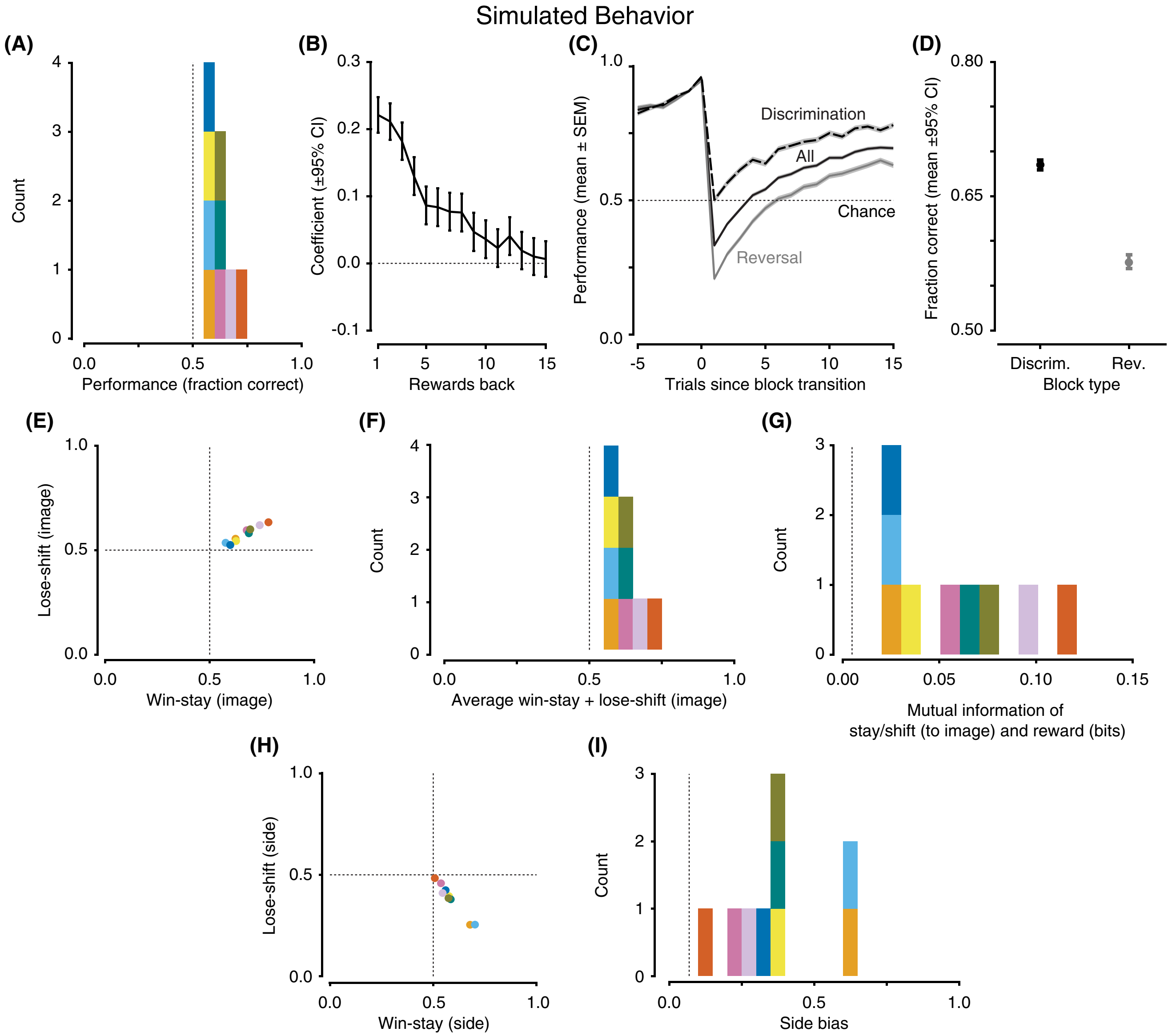
To obtain partially pooled parameter estimates (i.e., less noisy estimates, especially since α and β tend to compensate for one another (Ballard & McClure, 2019; Daw, 2011), we refit the best reinforcement learning model using a hierarchical Bayesian framework. We used MATLAB (Mathworks), the probabilistic programming language Stan (https://mc-stan.org), and the MATLAB interface, MatlabStan (https://mc-stan.org/users/interfaces/matlab-stan). We constructed hierarchical models separately for each monkey, with monkey-level parameters to govern session-level parameters for learning. Priors over monkey-level parameters, from which session-level means were drawn, were set as

$$\alpha \sim \mathrm{Beta}(1.2, 1.2)$$
$$\beta \sim \mathrm{Gamma}(4.82, 0.88)$$
$$b \sim \mathrm{Normal}(0, 1),$$

where the priors for α and β were taken from the literature (den Ouden et al., 2013; Gershman, 2016; Kanen et al., 2019). The gamma distribution was parameterized in terms of shape and scale. For all

Figure 4

Simulated Behavioral Features Demonstrate Squirrel Monkey-Like Reward Sensitivity and Side Bias



Note. (A) Distribution of performance was better than chance. (B) Logistic regression coefficients for choice as a function of reward history and side history shows dependence for many trials in the past. (C) Performance at block transitions for all blocks, and separately for Discrimination and Reversal blocks. Like actual performance, pretransition simulated data had no significant effect of Block Type which became significant after the transition. (D) Simulated performance was better in Discrimination blocks relative to Reversal blocks. (E) Image-based win–stay and lose–shift were both greater than 0.5. (F) The average win–stay + lose–shift was greater than 0.5. (G) The mutual information between stay/switch and reward on the previous trial is greater than random behavior. (H) Side-based win–stay and lose–shift demonstrates a side bias. (I) Side bias distribution. In panels A, E–I, each data point is the average for one monkey, across all sessions and blocks. Panels B–D are analysis of all the data, pooled across all monkeys, sessions, and blocks.

monkey-level variances, we used Cauchy$^+$(0, 1). For session-level α and β, we again used beta and gamma distributions, reparameterized in terms of mean and variance with parameters drawn from monkey-level distributions. Session-level bias was normally distributed, with mean and variance drawn from monkey-level distributions. Parameter estimates in Figure 6A–G and Figure S7 are posteriors over monkey-level means. Parameter values reported in Figure 6H–K are the means of the session-level posteriors.

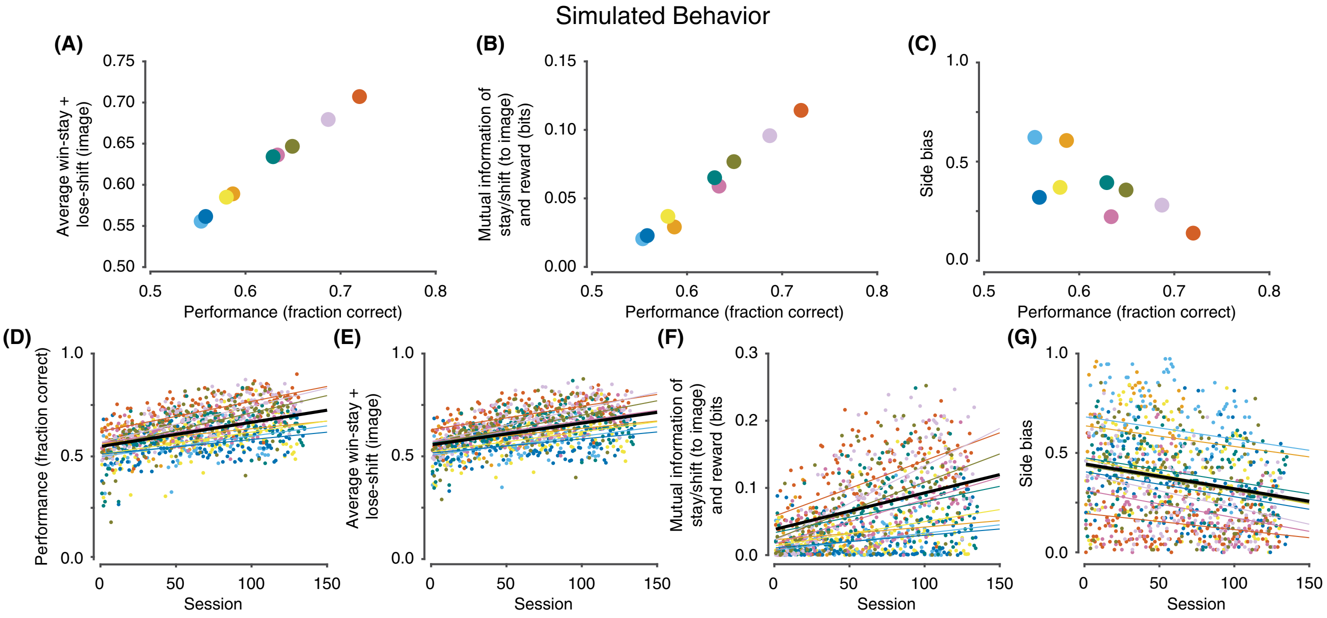
Results

Reward History and Side Bias Inform Strategy

We developed a deterministic reversal learning task in which chaired squirrel monkeys chose between two simultaneously presented images for delivery of milk reward. Images were presented in blocks of trials, and in each block, one image was associated with big reward and the other was associated with small reward. On each

Figure 5

Simulations Show Similar Relationships Between Simulated Performance, Reward Sensitivity, Side Bias, and Training



Note. (A) Average win–stay + lose–shift showed a positive relationship with performance. (B) Mutual information between stay/shift and reward as a function of performance. (C) Side bias as a function of performance. (D) Performance improved with training. (E) Average win–stay + lose–shift improved with training. (F) Mutual information between stay/switch and reward increased with training. (G) Side bias decreased with training. Black line shows the fixed effect and thin colored lines show individual monkey random effects. Colors depict individual monkeys and are consistent across figures. In panels A–C, each dot is the average for one monkey, across all sessions and blocks. In panels D–G, each dot is the average for one session, across all blocks.

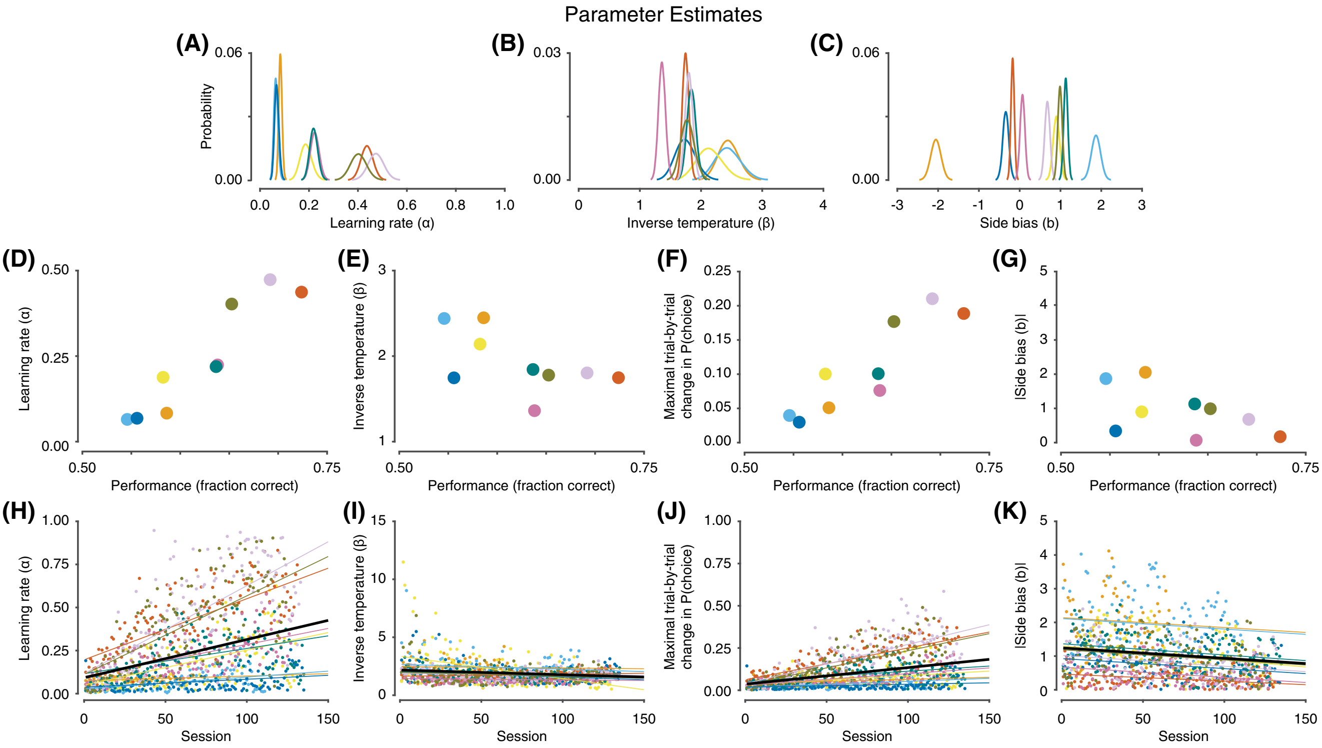
trial, monkeys were presented with two images, each on the left/right half of a touchscreen and physically touching an image yielded reward (Figure 1A). Selecting the big reward image (which we call the correct choice) for 80% of the past 15 trials triggered a block transition, uncued to the monkey. Blocks switched between Discrimination blocks and Reversal blocks (Figure 1B). At the beginning of each Discrimination block, two images were randomly selected from a large library of images and assigned to big/small reward. At the beginning of each Reversal block, the two images swapped reward contingencies. On average, sessions lasted for 146 (*SD* 14) trials and monkeys completed 4.8 (*SD* 1.9) blocks.

Assuming animals are sensitive to differences in reward volume and seek to maximize reward, the optimal strategy in this task is an image-based win–stay/lose–shift policy: select the same image if it yielded big reward on the previous trial, switch if it yielded small reward. After training, monkeys reliably switched their choices at block transitions, when contingencies switched (Figure 1C and Figure S4A). However, although animals performed significantly better than chance (Wilcoxon signed-rank test, $p < .01$), they performed worse than win–stay/lose–shift, an optimal strategy that would yield the large reward on ~96% of trials (Figure 2A). To understand how monkeys solved this task, we fit logistic regression models to predict choice as a function of reward history and side history. Unlike the optimal policy, which maintains a memory of reward on just the most recent trial, monkeys maintained a recency-weighted memory of reward history up to 10 trials in the past to inform choices (Figure 2B). We also found that monkeys were faster to transition from Reversal → Discrimination blocks than from Discrimination → Reversal blocks (Figure 2C, D). A two-way analysis of variance (ANOVA; Block Type × Trials) was performed separately for trials before and after the block transition. Before the block transition, there was no significant effect of Block Type ($F_{1,\ 96} = 1.32$, $p = .25$). After the transition, this effect became significant ($F_{1,\ 240} = 270.13$, $p < .0001$). Performance in Discrimination blocks was better than in Reversal blocks (*M* [95% CI] 0.685 [0.679, 0.690] fraction correct in Discrimination blocks, 0.576 [0.568, 0.583] fraction correct in Reversal blocks). This asymmetry in block performance is consistent with the idea that Reversal blocks, but not Discrimination blocks, require unlearning of previously learned associations in addition to learning new associations.

Following reward, monkeys can implement two distinctive strategies: choose to repeat choices to the same image, regardless of side (image-based win–stay) or repeat choices to the same side, regardless of image (side-based win–stay). To better define monkeys' behavioral strategy, we first quantified win–stay and lose–shift tendencies in image-based coordinates (Figure 2E). Each point indicates the fraction of trials in which monkeys stayed after receiving big reward (*x*-axis) and shifted after receiving the small reward (*y*-axis). Plotted this way, points in the top right and bottom left quadrants indicate reward-sensitive behavior (win–stay/lose–shift and win–shift/lose–stay, respectively) and points in the top left and bottom right quadrants indicate reward-insensitive behavior (shift and stay, respectively, regardless of reward). All monkeys were in the top right quadrant, indicating that they were demonstrating both win–stay and lose–shift behavior following outcomes

Figure 6

Relationship Between Model Parameters, Simulated Performance, and Training



Note. (A) Estimate learning rates for all monkeys. (B) Estimated inverse temperatures for all monkeys. (C) Estimated side biases for all monkeys. (D) As performance improved, the learning rate increased. (E) Inverse temperature showed no significant linear association with performance. (F) The maximal trial-by-trial change in *P*(choice), which partially accounts for the interaction of both the learning rate and the inverse temperature, increased as performance improved. (G) The absolute side bias showed no significant relationship with performance. (H) Learning rates improved with training. (I) The inverse temperature did not change throughout training. (J) The maximal trial-by-trial change in *P*(choice) increased with training. (K) Side bias decreased with training. Black line shows the fixed effect and thin colored lines show individual monkey random effects. Colors denote individual monkeys and are consistent between figures.

(Wilcoxon signed-rank test, $p < .01$ for each). We quantified reward sensitivity by computing the average win–stay + lose–shift per animal. This metric is 1.0 for perfect win–stay/lose–shift behavior, 0.5 for reward-insensitive behavior, and intermediate for reward-sensitive behavior. Consistent with prior analyses, animals demonstrated reward-sensitive behavior (Figure 2F; Wilcoxon signed-rank test, $p < .01$). However, one shortcoming of this metric is it places equal emphasis on win–stay and lose–shift. Because *P*(lose) is fairly low in this task, behavior in response to losses does not impact overall performance as strongly as response to reward. To work around this pitfall, we computed the mutual information between reward on the previous trial and stay/switch behavior on the current trial, as it accounts for the base rate of *P*(lose), Figure 2G. Perfect win–stay/lose–shift behavior results in 1 bit of information and reward-insensitive behavior results in 0 bits. Consistent with the average win–stay + lose–shift analysis, monkeys demonstrated reward-sensitive behavior (Wilcoxon signed-rank test, $p < .01$).

One notable behavioral suboptimality we observed was a bias toward a particular side (a rightward side bias can be seen in Figure S4A; Bari et al., 2019; Friedman et al., 2017). To better understand this side bias, we quantified win–stay/lose–shift in side-based coordinates (Figure 2H). Most monkeys fell in the bottom right quadrant, consistent with a reward-insensitive tendency to favor a particular side (win–stay >0.5 and lose–switch <0.5, Wilcoxon signed-rank test, $p < .01$ for each). We quantified side bias with a side bias metric (0 for uniformly split choices, 1 for exclusive choice of one side) and observed a wide distribution, indicating an average tendency for side-biased behavior (Figure 2I; Wilcoxon signed-rank test, $p < .01$). Individual session data showed a similar trend (Figure S1).

Taken together, these results argue that monkeys solve this task by integrating reward over many trials to inform choices, and that this strategy is corrupted by a side bias.

Squirrel Monkeys Develop Reward Sensitivity and Reduce Side Bias With Training

The wide range of performance allowed us to relate behavioral performance to the behavioral metrics we defined. First, we found that the average win–stay + lose–shift increased with better performance (linear slope 0.82, $t_7 = 28.23$, $p < .0001$), consistent with the optimality of image-based win–stay/lose–shift (Figure 3A). Similarly, the mutual information between reward and stay/shift increased with performance (Figure 3B; linear slope 0.50, $t_7 = 19.54, p < .0001$). Next, we focused on the side bias (Figure 3 C). We found that for poor performance, side bias was generally high and decreased with improved performance (mean: linear slope −1.98, $t_7 = -3.17, p = .016$).

The large number of sessions per monkey additionally allowed us to quantify the effects of training. First, we found that performance increased with the number of sessions (Figure 3D; linear slope 1.3×10^{-3}, $t_{1091} = 7.85$, $p < .0001$). Similarly, we found that mean block lengths decreased with training (Figure S2; linear slope -6.8×10^{-2}, $t_{1064} = -4.72$, $p < .0001$). This was partly due to increased reward sensitivity to images. The average win–stay + lose–shift increased with training (Figure 3E; linear slope 1.1×10^{-3}, $t_{1091} = 11.04$, $p < .0001$).

Similarly, the mutual information between reward and stay/switch increased (Figure 3F; linear slope 4.86×10^{-4}, $t_{1091} = 5.58$, $p < .0001$). The improvement in performance was also partly due to a decrease in side bias. The side bias decreased with training (Figure 3G; linear slope -1.5×10^{-3}, $t_{1091} = -8.51$, $p < .0001$). Individual session data showed a similar trend (Figure S3). In summary, squirrel monkeys improve with training, partly due to increased reward sensitivity to images, and partly due to a decrease in side bias.

Reinforcement Learning Modeling Captures Key Features of Behavior

Since we found that squirrel monkeys integrated rewards over many trials, we adapted the Rescorla–Wagner model, a commonly used model in reinforcement learning (Rescorla, 1972). This model maintains a running estimate of the values of images and chooses based on the relative values of the presented images. Image values are learned by recency-weighted reward history, which allows the model to adapt behavior flexibly when reward contingencies change. We considered a number of model variants: equivalent versus differential learning from better-than-expected and worse-than-expected outcomes, forgetting of unchosen image values, learning the values of actions (e.g., if leftward choices were recently rewarded, then increase probability of leftward choices), mixtures of reinforcement learning and win–stay/lose–shift strategies, and nuisance parameters, like side bias and choice autocorrelation. We fit individual sessions using maximum likelihood estimation and selected the best model using Bayesian information criteria, which selects the best fit model while penalizing overly complex models. The best model was among the simplest—learning of image values with equivalent learning from better/worse outcomes and a side bias mechanism (Table S1). Importantly, this model was strongly preferred over noise models (which include nuisance parameters but no learning of image values), suggesting that learning image values was consistent with real behavior. Armed with a simple and tractable model, we sought to determine how well it described real behavior.

First, we observed that the model fit behavioral data well (data not shown). However, model fits run the risk of overfitting to data (Palminteri et al., 2017). A stronger approach is to take advantage of the generative modeling framework: simulate fictive data and assess how well simulated data matches real behavioral data. Visually, we observed a qualitative correspondence between raw behavior and simulations (Figure S4B). Across all simulated monkeys, we observed that simulated behavior performed better than chance, similar to real behavior (Figure 4A; Wilcoxon signed-rank test, $p < .01$). Simulated behavior exhibited a dependence on reward history for many trials into the past (Figure 4B). Like real squirrel monkeys, simulated monkeys were faster to transition to new Discrimination blocks than to new Reversal blocks (Figure 4C). There was no significant effect of Block Type prior to block transitions ($F_{1,\,96} = 0.50$, $p = .48$), which became significant after the transition ($F_{1,\,240} = 274.78$, $p < .0001$). Performance for Discrimination and Reversal blocks were comparable to real behavior (Figure 4D; *M* [95% CI], Discrimination: 0.685 [0.679, 0.691], Reversal: 0.576 [0.568, 0.584]).

Simulated behavior exhibited features of reward sensitivity to images, with image-based win–stay and lose–shift both >0.5 (Figure 4E; Wilcoxon signed-rank test, $p < .01$ for each). The average win–stay + lose–shift was >0.5, indicating reward-sensitive behavior (Wilcoxon signed-rank test, $p < .01$) and mutual information between reward and stay/switch was similarly skewed away from 0 bits, indicating reward sensitivity (Wilcoxon signed-rank test, $p < .01$; Figure 4F, G). Simulated behavior also exhibited suboptimal features of side bias (Figure 4H; win–stay >0.5 and lose–switch <0.5, $p < .01$ for each). Side bias was similarly wide (Figure 4I). On a monkey-by-monkey basis, there was a strong correspondence between each of these metrics for real and simulated data (Figure S5).

We addressed the relationship between simulated behavioral metrics and performance. We observed a strong linear dependence between performance and average win–stay + lose–shift (Figure 5 A; linear slope 0.91, $t_7 = 47.85$, $p < .0001$). There was likewise a strong association between performance and mutual information between reward and stay/switch (Figure 5B; linear slope 0.57, $t_7 = 23.63$, $p < .0001$). Side bias decreased with performance (Figure 5C; linear slope −1.99, $t_7 = -3.06$, $p = .18$).

We also addressed the relationship between behavioral metrics and training. Simulated behavior exhibited an increase in performance with training (Figure 5D; linear slope 1.12×10^{-3}, $t_{1091} = 8.00$, $p < .0001$). The average win–stay + lose–shift improved with training (Figure 5E; linear slope 1.05×10^{-3}, $t_{1091} = 8.56$, $p < .0001$) and the mutual information between reward and stay/switch improved with training (Figure 5F; linear slope 5.44×10^{-4}, $t_{1091} = 4.65$, $p < .0001$). Side bias decreased with training (Figure 5G; linear slope -1.25×10^{-3}, $t_{1091} = -6.35$, $p < .0001$).

Taken together, these results argue that our reinforcement learning model with two core features—learning of image values and a side bias—is sufficient to capture key features of real behavior.

Model Parameters Provide Interpretable Insight Into Behavioral Strategy

A key advantage of our generative modeling approach, beyond traditional summary statistics (e.g., mean performance, block lengths, etc.), is the ability to provide intuitive explanations for how behavior was generated. Our model has two key components—a learning component and a decision component. The learning component determines the image values and the decision component turns the relative image values into a decision. The model has three parameters, which we detail below: learning rate (α), inverse temperature (β), and side bias (Figure S6A).

The learning rate, which affects the learning component, determines how quickly image values are updated following an outcome (Figure S6B). At its extremes, a learning rate closer to 1 means learning from only the most recent trials and a learning rate closer to 0 means learning from many previous trials. In this task, higher learning rates are adaptive, and correspond with faster block transitions and better performance. The inverse temperature determines choice stochasticity, or how deterministically the model acts (Figure S6C). High values of inverse temperature correspond to more deterministic choice functions—the agent will opt to choose the image with a higher value, even if the difference is small. Small values of inverse temperature correspond to more random behavior—the agent will still choose the image with lower value with reasonable probability. In this task, there is a more complex correspondence between inverse temperature and performance. High values of inverse temperature correspond to behavior that better maximizes reward when the better option is known, but tends to perseverate at block transitions. In general, higher values of inverse temperature correspond with better performance. The side bias determines the model's preference for a stimulus location, regardless of relative image values (Figure S3D). Nonzero values of side bias are strictly maladaptive in this task and correspond to poorer performance.

To better estimate model parameters, we adopted a hierarchical Bayesian strategy to fit the reinforcement learning model and obtain monkey- and session-level parameter estimates (Figure 6A–C and Figure S7). We related these parameter estimates to performance to gain better insight. We found that the learning rate improved with increased performance (Figure 6D; linear slope 2.47, $t_7 = 7.70$, $p < .0001$). In contrast, the inverse temperature showed no significant linear association with performance (Figure 6E; linear slope −3.23, $t_7 = -2.00$, $p = .086$). Because changes in learning rates and inverse temperatures can partially compensate for one another (small increase in learning rate can be compensated for by a small decrease in inverse temperature; Daw, 2011), we sought to measure their combined effect on P(choice). The maximal trial-by-trial change in P(choice), which partially accounts for this interaction, showed an increase with performance (Figure 6F; linear slope 1.00, $t_7 = 5.99$, $p = 5.5 \times 10^{-4}$). Finally, the absolute value of side bias showed no change with performance (Figure 6G; linear slope −5.91, $t_7 = -1.80$, $p = .11$; see Discussion).

We next sought to estimate how these parameters changed with training. Learning rates increased with training, which yields better performance (Figure 6H; linear slope 2.22×10^{-3}, $t_{1091} = 3.78$, $p = 1.7 \times 10^{-4}$). In contrast, inverse temperature showed no significant change with training (Figure 6I; linear slope -4.02×10^{-3}, $t_{1091} = -1.86$, $p = .06$). This is noteworthy since animals consistently adopted suboptimal inverse temperatures and would have benefited from increased β values (Figure S8; see Discussion). The maximal trial-by-trial change in P(choice) improved with training (Figure 6J; linear slope 9.79×10^{-4}, $t_{1091} = 3.53$, $p = 4.7 \times 10^{-4}$).

Finally, side bias decreased with training, which permitted better performance (Figure 6K; linear slope -3.10×10^{-3}, $t_{1091} = -5.62$, $p < .0001$). We obtained similar results after within-animal normalization by z-scoring, though with a small decrease in inverse temperature with training (Figure S9).

In summary, the reinforcement modeling approach provides a simple and compelling model for understanding how squirrel monkeys solve a reversal learning task and provides a tool for interpreting how the inner mechanisms relate to performance and how they change with training.

Discussion

The power of comparative neuroscience to dissect cognition relies on the use of behavioral tasks that engage similar cognitive mechanisms in different species. Here, we show that squirrel monkeys solve a reversal learning task, a frequently used behavioral paradigm, similarly to other species. We found that these animals integrate reward history over many trials to dictate choices, a commonly observed reinforcement learning motif across species.

Using generative modeling, we explicitly tested a number of hypotheses about the strategies animals applied to harvest reward. We found that animal behavior was consistent with a remarkably simple strategy: reward history integration over many trials (~5–10) and a bias for a particular side. This model outperformed a number of other reasonable hypotheses. In particular, we found that a one learning rate model outperformed models with two learning rates. This is notable since models with two learning rates, which allow for separate learning from positive and negative reward prediction errors, are commonly found to better explain behavioral data (Averbeck, 2017; Dorfman et al., 2019; Frank et al., 2004; Gershman, 2015; Grossman et al., 2020; Niv et al., 2012; Taswell et al., 2018), although these tasks often have different reward statistics than what we used here.

We found that animals did not implement a pure or noisy win–stay/lose–shift strategy, either in isolation or mixed with a reinforcement learning strategy. Why did not animals approximate the optimal strategy in this task? We emphasize that this strategy is only optimal if animals are sensitive to differences in reward volume and seek to maximize reward. Although win–stay/lose–shift is optimal on this particular task variant, it may not be adaptive across task variants in general. Reward probabilities vary drastically and dynamically in natural environments and, presumably, by integrating reward history, animals would continue to perform well if reward probabilities changed. Reward could be optimized by tweaking parameters (e.g., adjusting learning rates), rather than changing the entire behavioral strategy (Doya, 2002). Another potential reason is that incorrect choices still yielded reward, allowing animals to perform well enough despite using a suboptimal strategy. Different neural structures are engaged when animals gain from both options than when they can lose from selecting the incorrect option (Taswell et al., 2018). It is possible that animals would have better approximated win–stay/lose–shift if incorrect choices yielded no reward or mild punishment.

Squirrel monkeys did not adopt optimal combinations of learning rates and inverse temperatures to maximize reward (Figure S8). Although animals had some exposure to the basics of the task (data not included), early in training, we would not expect animals to implement optimal parameter combinations. With training, they may approximate ideal parameter combinations with greater knowledge of task statistics. Indeed, we found that learning rates increased with training, which allows for better performance. Inverse temperatures, however, did not increase, which would be expected to optimize reward. Lower inverse temperatures meant squirrel monkeys made choices more randomly. This finding is consistent with the notion that squirrel monkeys maintained a high level of exploratory behavior, which may be a ubiquitous feature of behavior even when task demands encourage more deterministic choice behavior (Ebitz et al., 2019; Pisupati et al., 2021). Limited attention may also contribute to suboptimal exploratory behavior.

There are at least two kinds of "learning" monkeys implement in this task. On the trial-by-trial timescale, monkeys learn the image associations within a block, which is captured by our reinforcement learning models. On the session-to-session timescale, monkeys demonstrate a form of "metalearning," in which they adapt to the demands of the task structure, which gives rise to the changes in parameters over sessions (Figure 6). Our modeling approach is able to capture learning in both of these domains, which we believe offers a unique strength when interpreting manipulations. For example, monkeys may only show a "meta-learning" deficit following drug manipulation, but may otherwise be capable of trial-by-trial reinforcement learning.

Side-specific biases were a key feature of behavior in this task. These low-level idiosyncratic tendencies are often ubiquitous features of animal behavior that can be trained out with longer training protocols, as is often done with macaques (Lau & Glimcher, 2005; Tsutsui et al., 2016). In rodents, many studies have noted spatial biases in well-trained mice and rats (Bari et al., 2019; Grossman et al., 2020; Miller et al., 2017). Despite the presence of suboptimal side biases, these studies have helped elucidate the neural structures and representations underlying reinforcement learning. This is because behavioral strategies can be thought of as a mixture of a cognitive, reward-history-dependent component, and an idiosyncratic low-level bias. By properly modeling behavior, the effect of a side bias can be accounted for, allowing for interpretation of the reinforcement learning component. This was in fact our rationale for testing a large number of models with a number of "nuisance" parameters (bias and perseveration)—they not only fit the data better but allow for better estimation of parameters and latent variables. We additionally view the presence of side bias to be interesting in its own right. It can be thought of as a simple heuristic that allows for reward accumulation with little cognitive cost. We believe the presence of a mild-to-moderate level of side bias provides the dynamic range to observe both improvements and impairments in performance as a function of manipulation. These side biases are likely not controlled by the same brain regions that engender flexible behavior (Balleine & O'Doherty, 2010) and can be exaggerated by shutting down these structures (Bari et al., 2019). Because they are independent, it is conceivable that manipulation may improve performance by reducing the reliance on side bias, an effect that would be missed if animals were overtrained. In our study, the reduction in side bias and increase in learning rate were correlated during training (Figure 6H–K), not because the two processes rely on the same brain structures but likely because reduced side bias and increased learning rate yield better performance.

One potentially inconsistent finding is that the side bias metric decreased with performance (Figures 3C and 5C), while the bias parameter from the reinforcement learning model did not show a statistically significant change (Figure 6G). This is likely because this particular analysis was underpowered. When we analyzed data at the session level, we observed a significant decrease in both the mean and variance of the bias parameter (Figure S3C). For poor performance, side bias was highly variable. This is because poor performance could be the result of a strong bias to one side, or it could be the result of random, reward-insensitive behavior with no side bias. With better performance, side bias decreased in mean and variance, because a strong bias places an upper bound on performance, no matter how reward sensitive animals are.

Generative modeling allows us to test hypotheses that may be beyond the reach of simple summary statistics. For example, it is reasonable that animals could have computed action values in addition to a side bias (e.g., in a task variant where computing action values may be adaptive). It is not clear how the choice-based win–stay/lose–shift analyses we used (Figure 2H), which can test whether animals implement reward sensitivity to choices versus side bias, would help if animals implemented a mixture of the two. With the generative modeling approach, as long as the model is recoverable, then this hypothesis would be simple to test (Wilson & Collins, 2019).

Our modeling approach, while generally successful, did not perfectly recapitulate all behavioral features. One notable failure was the inability to capture the slight recovery of performance in the one trial immediately after a Reversal block began (compare Figures 2C and 4C). Interestingly, we found that simulated data with a mixture of reinforcement learning and win–stay/lose–shift was able to partially recapitulate this phenotype. However, the fact that none of these models fit animal behavior well (Table S1) argues that win–stay/lose–shift is not a cardinal feature of behavior, at least given our model selection pipeline. Interestingly, win–stay/lose–shift may only be a strategy animals implement on particular trials (Iigaya et al., 2018), which may disfavor a model that assumes win–stay/lose–shift is implemented on every trial. Perhaps squirrel monkeys implement win–stay/lose–shift only following large magnitude negative reward prediction errors, accounting for behavior in the trials immediately after a block change, and otherwise implement reinforcement learning. Learning rates might also change as a function of recent reward statistics, yielding nonstationary behavioral strategies (Behrens et al., 2007; Grossman et al., 2020; Nassar et al., 2012).

One strength of generative modeling is that it allows for interpretable insights into manipulations, particularly across species. Parameter estimates (Figure S7) may be compared across groups to gain insight into the effects of disease or manipulations (Aylward et al., 2019; Huys et al., 2013; Kanen et al., 2021). A complementary approach is to extract the latent variables governed by these parameters and correlate them with neural activity (Bari et al., 2019; Findling et al., 2019; Samejima et al., 2005). Insights at the level of parameters or latent variables may aid the development of novel therapies, since the development pipeline for nervous system therapeutics often stalls due to lack of objective biomarkers of success (Kola, 2008; Paulus et al., 2016). Since these types of models have theoretical underpinnings, parameter changes may be interpreted through the lens of theory. For example, the volatility of the environment should modulate learning rates (Behrens et al., 2007), beliefs about the causal structure of the environment should modulate asymmetric learning from good and bad outcomes (Dorfman et al., 2019), and the complexity of action space should govern the inverse temperature and perseveration (Gershman, 2020).

Our study builds on the notion that model-free reinforcement learning is a common computational motif present across animal models. The reinforcement learning model we chose is a fairly general algorithm that is not specific to the task the squirrel monkeys performed. In fact, to best study the cognitive mechanisms underlying this algorithm across species, we may need to adjust the task across species to account for species-specific intelligence. Humans performing a deterministic reversal learning task would almost certainly discover that win–stay/lose–shift was the optimal policy and exploit it. It is well known that humans search for structure even in purely random data (Kahneman & Tversky, 1972), an inductive bias that would be leveraged to discover win–stay/lose–shift as an optimal strategy. Another feature of the task that can be exploited is the rule governing block switches—80% correct in the past 20 trials. If animals track this, they can guess block transitions. Rhesus macaques overtrained on a deterministic reversal learning paradigm eventually learn expected block lengths (Jang et al., 2015). Therefore, to best study model-free reinforcement learning in humans and macaques may require a probabilistic reversal learning task without overtraining (to avoid win–stay/lose–shift policies), or a task where the probabilities drift slowly across time, without clear reversals (to avoid learning expected block lengths; Daw et al., 2006). An apparent advantage of squirrel monkeys is the ability to train on a simple reversal learning without needing to control for sophisticated strategies, although behavior on different variants of the task would need to be compared to verify this intuition.

Squirrel monkeys offer unique advantages relative to other model organisms. Compared to macaques, squirrel monkeys are small, easy to manage and house in large numbers, and more cost-effective per subject. Marmosets, which may offer similar advantages, typically do not complete many within-session reversals and often require several days to learn new associations (Clarke et al., 2011; Takemoto et al., 2015), similar to rodents (Izquierdo et al., 2017). This means that well-trained squirrel monkeys may more readily approximate human strategies, which would significantly aid the ability to translate insights. Our results highlight the utility of reinforcement learning modeling and validate squirrel monkeys as a useful behavioral neuroscience model.

References

Abee, C. R. (2000). Squirrel monkey (Saimiri spp.) research and resources. *ILAR Journal*, *41*(1), 2–9. https://doi.org/10.1093/ilar.41.1.2

Averbeck, B. B. (2017). *Amygdala and ventral striatum population codes implement multiple learning rates for reinforcement learning* [Conference session]. 2017 IEEE Symposium Series on Computational Intelligence (SSCI) Proceedings, Piscataway, New Jersey, United States. https://doi.org/10.1109/SSCI.2017.8285354

Aylward, J., Valton, V., Ahn, W. Y., Bond, R. L., Dayan, P., Roiser, J. P., & Robinson, O. J. (2019). Altered learning under uncertainty in unmedicated mood and anxiety disorders. *Nature Human Behaviour*, *3*(10), 1116–1123. https://doi.org/10.1038/s41562-019-0628-0

Ballard, I. C., & McClure, S. M. (2019). Joint modeling of reaction times and choice improves parameter identifiability in reinforcement learning models. *Journal of Neuroscience Methods*, *317*, 37–44. https://doi.org/10.1016/j.jneumeth.2019.01.006

Balleine, B. W., & O'Doherty, J. P. (2010). Human and rodent homologies in action control: Corticostriatal determinants of goal-directed and habitual action. *Neuropsychopharmacology*, *35*(1), 48–69. https://doi.org/10.1038/npp.2009.131

Bari, B. A., Grossman, C. D., Lubin, E. E., Rajagopalan, A. E., Cressy, J. I., & Cohen, J. Y. (2019). Stable representations of decision variables for flexible behavior. *Neuron*, *103*(5), 922–933. Article e7. https://doi.org/10.1016/j.neuron.2019.06.001

Behrens, T. E. J., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. S. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience*, *10*(9), 1214–1221. https://doi.org/10.1038/nn1954

Boinski, S. (1999). The social organizations of squirrel monkeys: Implications for ecological models of social evolution. *Evolutionary Anthropology: Issues, News, and Reviews*, *8*(3), 101–112. https://doi.org/10.1002/(SICI)1520-6505(1999)8:3<101::AID-EVAN5>3.0.CO;2-O

Brigman, J. L., Ihne, J., Saksida, L. M., Bussey, T. J., & Holmes, A. (2009). Effects of subchronic phencyclidine (PCP) treatment on social behaviors, and operant discrimination and reversal learning in C57BL/6J mice. *Frontiers in Behavioral Neuroscience*, *3*. Article 2. https://doi.org/10.3389/neuro.08.002.2009

Clarke, H. F., Hill, G. J., Robbins, T. W., & Roberts, A. C. (2011). Dopamine, but not serotonin, regulates reversal learning in the marmoset caudate nucleus. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *31*(11), 4290–4297. https://doi.org/10.1523/JNEUROSCI.5066-10.2011

Costa, V. D., Dal Monte, O., Lucas, D. R., Murray, E. A., & Averbeck, B. B. (2016). Amygdala and ventral striatum make distinct contributions to reinforcement learning. *Neuron*, *92*(2), 505–517. https://doi.org/10.1016/j.neuron.2016.09.025

Crawley, J. N. (2008). Behavioral phenotyping strategies for mutant mice. *Neuron*, *57*(6), 809–818. https://doi.org/10.1016/j.neuron.2008.03.001

Daw, N. D. (2011). Trial-by-trial data analysis using computational models. In M. R. Delgado, E. A. Phelps, & T. W. Robbins (Eds.), *Decision making, affect, and learning: Attention and performance XXIII* (p. 23). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199600434.003.0001

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

Dawson, G. R., & Tricklebank, M. D. (1995). Use of the elevated plus maze in the search for novel anxiolytic agents. *Trends in Pharmacological Sciences*, *16*(2), 33–36. https://doi.org/10.1016/S0165-6147(00)88973-7

den Ouden, H. E. M., Daw, N. D., Fernandez, G., Elshout, J. A., Rijpkema, M., Hoogman, M., Franke, B., & Cools, R. (2013). Dissociable effects of dopamine and serotonin on reversal learning. *Neuron*, *80*(4), 1090–1100. https://doi.org/10.1016/j.neuron.2013.08.030

Dorfman, H. M., Bhui, R., Hughes, B. L., & Gershman, S. J. (2019). Causal inference about good and bad outcomes. *Psychological Science*, *30*(4), 516–525. https://doi.org/10.1177/0956797619828724

Doya, K. (2002). Metalearning and neuromodulation. *Neural Networks*, *15*(4–6), 495–506. https://doi.org/10.1016/S0893-6080(02)00044-8

Ebitz, R. B., Sleezer, B. J., Jedema, H. P., Bradberry, C. W., & Hayden, B. Y. (2019). Tonic exploration governs both flexibility and lapses. *PLoS Computational Biology*, *15*(11). Article e1007475. https://doi.org/10.1371/journal.pcbi.1007475

Ersche, K. D., Roiser, J. P., Abbott, S., Craig, K. J., Müller, U., Suckling, J., Ooi, C., Shabbir, S. S., Clark, L., Sahakian, B. J., Fineberg, N. A., Merlo-Pich, E. V., Robbins, T. W., & Bullmore, E. T. (2011). Response perseveration in stimulant dependence is associated with striatal dysfunction and can be ameliorated by a D(2/3) receptor agonist. *Biological Psychiatry*, *70*(8), 754–762. https://doi.org/10.1016/j.biopsych.2011.06.033

Fenton, W. S., Stover, E. L., & Insel, T. R. (2003). Breaking the log-jam in treatment development for cognition in schizophrenia: NIMH perspective. *Psychopharmacology*, *169*(3–4), 365–366. https://doi.org/10.1007/s00213-003-1564-1

Findling, C., Skvortsova, V., Dromnelle, R., Palminteri, S., & Wyart, V. (2019). Computational noise in reward-guided learning drives behavioral variability in volatile environments. *Nature Neuroscience*, *22*(12), 2066–2077. https://doi.org/10.1038/s41593-019-0518-9

Flint, J., & Shifman, S. (2008). Animal models of psychiatric disease. *Current Opinion in Genetics & Development*, *18*(3), 235–240. https://doi.org/10.1016/j.gde.2008.07.002

Frank, M. J., Samanta, J., Moustafa, A. A., & Sherman, S. J. (2007). Hold your horses: impulsivity, deep brain stimulation, and medication in parkinsonism. *Science 318*(5854), 1309–1312. https://doi.org/10.1126/science.1146157

Frank, M. J., Seeberger, L. C., & O'Reilly, R. C. (2004). By carrot or by stick: Cognitive reinforcement learning in parkinsonism. *Science*, *306*(5703), 1940–1943. https://doi.org/10.1126/science.1102941

Friedman, A., Homma, D., Bloem, B., Gibb, L. G., Amemori, K.-I., Hu, D., Delcasso, S., Truong, T. F., Yang, J., Hood, A. S., Mikofalvy, K. A., Beck, D. W., Nguyen, N., Nelson, E. D., Toro Arana, S. E., Vorder Bruegge, R. H., Goosens, K. A., & Graybiel, A. M. (2017). Chronic stress alters striosome-circuit dynamics, leading to aberrant decision-making. *Cell*, *171*(5), 1191–1205. Article e28. https://doi.org/10.1016/j.cell.2017.10.017

Gershman, S. J. (2015). Do learning rates adapt to the distribution of rewards? *Psychonomic Bulletin & Review*, *22*(5), 1320–1327. https://doi.org/10.3758/s13423-014-0790-3

Gershman, S. J. (2016). Empirical priors for reinforcement learning models. *Journal of Mathematical Psychology*, *71*, 1–6. https://doi.org/10.1016/j.jmp.2016.01.006

Gershman, S. J. (2020). Origin of perseveration in the trade-off between reward and complexity. *Cognition*, *204*, Article 104394. https://doi.org/10.1016/j.cognition.2020.104394

Gershman, S. J., & Lai, L. (2021). The reward-complexity trade-off in schizophrenia. *Computational Psychiatry*, *5*(1), 38–53. https://doi.org/10.5334/cpsy.71

Grossman, C. D., Bari, B. A., Cohen, J. Y. (2020). Serotonin neurons modulate learning rate through uncertainty. *bioRxiv*. https://doi.org/10.1101/2020.10.24.353508

Harris, C., Aguirre, C. G., Kolli, S., Das, K., Izquierdo, A., Soltani, A. (2020). Unique features of stimulus-based probabilistic reversal learning. *bioRxiv*. https://doi.org/10.1101/2020.09.24.310771

Heathcote, A., Brown, S. D., & Wagenmakers, E. J. (2015). An introduction to good practices in cognitive modeling. In B. U. Forstmann & E.-J. Wagenmakers (Eds.), *An introduction to model-based cognitive neuroscience* (pp. 25–48). Springer. https://doi.org/10.1007/978-1-4939-2236-9_2

Huys, Q. J. M., Maia, T. V., & Frank, M. J. (2016). Computational psychiatry as a bridge from neuroscience to clinical applications. *Nature Neuroscience*, *19*(3), 404–413. https://doi.org/10.1038/nn.4238

Huys, Q. J., Pizzagalli, D. A., Bogdan, R., & Dayan, P. (2013). Mapping anhedonia onto reinforcement learning: A behavioural meta-analysis. *Biology of Mood & Anxiety Disorders*, *3*(1), Article 12. https://doi.org/10.1186/2045-5380-3-12

Iigaya, K., Fonseca, M. S., Murakami, M. Mainen, Z. F., & Dayan, P. (2018). An effect of serotonergic stimulation on learning rates for rewards apparent after long intertrial intervals. *Nature Communications*, *9*(1), Article 2477. https://doi.org/10.1038/s41467-018-04840-2

Izquierdo, A., Brigman, J. L., Radke, A. K., Rudebeck, P. H., & Holmes, A. (2017). The neural basis of reversal learning: An updated perspective. *Neuroscience*, *345*, 12–26. https://doi.org/10.1016/j.neuroscience.2016.03.021

Izquierdo, A., & Jentsch, J. D. (2012). Reversal learning as a measure of impulsive and compulsive behavior in addictions. *Psychopharmacology*, *219*(2), 607–620. https://doi.org/10.1007/s00213-011-2579-7

Jang, A. I., Costa, V. D., Rudebeck, P. H., Chudasama, Y., Murray, E. A., & Averbeck, B. B. (2015). The role of frontal cortical and medial-temporal lobe brain areas in learning a Bayesian prior belief on reversals. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *35*(33), 11751–11760. https://doi.org/10.1523/JNEUROSCI.1594-15.2015

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, *3*(3), 430–454. https://doi.org/10.1016/0010-0285(72)90016-3

Kaiser, T., & Feng, G. (2015). Modeling psychiatric disorders for developing effective treatments. *Nature Medicine*, *21*(9), 979–988. https://doi.org/10.1038/nm.3935

Kanen, J. W., Ersche, K. D., Fineberg, N. A., Robbins, T. W., & Cardinal, R. N. (2019). Computational modelling reveals contrasting effects on reinforcement learning and cognitive flexibility in stimulant use disorder and obsessive-compulsive disorder: Remediating effects of dopaminergic D2/3 receptor agents. *Psychopharmacology*, *236*(8), 2337–2358. https://doi.org/10.1007/s00213-019-05325-w

Kanen, J. W., Luo, Q., Kandroodi, M. R., Cardinal, R. N., Robbins, T. W., Carhart-Harris, R. L., & den Ouden, H. E. (2021). Effect of lysergic acid diethylamide (LSD) on reinforcement learning in humans. *bioRxiv*. https://doi.org/10.1101/2020.12.04.412189

Katahira, K. (2015). The relation between reinforcement learning parameters and the influence of reinforcement history on choice behavior. *Journal of Mathematical Psychology*, *66*, 59–69. https://doi.org/10.1016/j.jmp.2015.03.006

Kelleher, R. T., Gill, C. A., Riddle, W. C., & Cook, L. (1963). On the use of the squirrel monkey in behavioral and pharmacological experiments. *Journal of the Experimental Analysis of Behavior*, *6*(2), 249–252. https://doi.org/10.1901/jeab.1963.6-249

Kola, I. (2008). The state of innovation in drug development. *Clinical Pharmacology and Therapeutics*, *83*(2), 227–230. https://doi.org/10.1038/sj.clpt.6100479

Lau, B., & Glimcher, P. W. (2005). Dynamic response-by-response models of matching behavior in rhesus monkeys. *Journal of the Experimental Analysis of Behavior*, *84*(3), 555–579. https://doi.org/10.1901/jeab.2005.110-04

Leeson, V. C., Robbins, T. W., Matheson, E., Hutton, S. B., Ron, M. A., Barnes, T. R., & Joyce, E. M. (2009). Discrimination learning, reversal, and set-shifting in first-episode schizophrenia: Stability over six years and specific associations with medication type and disorganization syndrome. *Biological Psychiatry*, *66*(6), 586–593. https://doi.org/10.1016/j.biopsych.2009.05.016

Markou, A., Chiamulera, C., Geyer, M. A., Tricklebank, M., & Steckler, T. (2009). Removing obstacles in neuroscience drug discovery: The future path for animal models. *Neuropsychopharmacology*, *34*(1), 74–89. https://doi.org/10.1038/npp.2008.173

Mason, L., Eldar, E., & Rutledge, R. B. (2017). Mood instability and reward dysregulation—A neurocomputational model of bipolar disorder. *JAMA Psychiatry*, *74*(12), 1275–1276. https://doi.org/10.1001/jamapsychiatry.2017.3163

Metha, J. A., Brian, M. L., Oberrauch, S., Barnes, S. A., Featherby, T. J., Bossaerts, P., Murawski, C., Hoyer, D., & Jacobson, L. H. (2020). Separating probability and reversal learning in a novel Probabilistic Reversal Learning task for mice. *Frontiers in Behavioral Neuroscience*, *13*, Article 270. https://doi.org/10.3389/fnbeh.2019.00270

Michely, J., Eldar, E., Martin, I. M., & Dolan, R. J. (2020). A mechanistic account of serotonin's impact on mood. *Nature Communications*, *11*(1), Article 2335. https://doi.org/10.1038/s41467-020-16090-2

Miller, K. J., Botvinick, M. M., & Brody, C. D. (2017). Dorsal hippocampus contributes to model-based planning. *Nature Neuroscience*, *20*(9), 1269–1276. https://doi.org/10.1038/nn.4613

Nassar, M. R., Rumsey, K. M., Wilson, R. C., Parikh, K., Heasly, B., & Gold, J. I. (2012). Rational regulation of learning dynamics by pupil-linked arousal systems. *Nature Neuroscience*, *15*(7), 1040–1046. https://doi.org/10.1038/nn.3130

Nestler, E. J., Gould, E., Manji, H., Buncan, M., Duman, R. S., Greshenfeld, H. K., Hen, R., Koester, S., Lederhendler, I., Meaney, M., Robbins, T., Winsky, L., & Zalcman, S. (2002). Preclinical models: Status of basic research in depression. *Biological Psychiatry*, *52*(6), 503–528. https://doi.org/10.1016/S0006-3223(02)01405-1

Niv, Y., Edlund, J. A., Dayan, P., & O'Doherty, J. P. (2012). Neural prediction errors reveal a risk-sensitive reinforcement-learning process in the human brain. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *32*(2), 551–562. https://doi.org/10.1523/JNEUROSCI.5498-10.2012

O'Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K., & Dolan, R. J. (2004). Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science*, *304*(5669), 452–454. https://doi.org/10.1126/science.1094285

Ottenheimer, D. J., Bari, B. A., Sutlief, E., Fraser, K. M., Kim, T. H., Richard, J. M., Cohen, J. Y., & Janak, P. H. (2020). A quantitative reward prediction error signal in the ventral pallidum. *Nature Neuroscience*, *23*(10), 1267–1276. https://doi.org/10.1038/s41593-020-0688-5

O'Shea, D. J., Kalanithi, P., Ferenczi, E. A., Hsueh, B., Chandrasekaran, C., Goo, W., Diester, I., Ramakrishnan, C., Kaufman, M. T., Ryu, S. I., Yeom, K. W., Deisseroth, K., & Shenoy, K. V. (2018). Development of an optogenetic toolkit for neural circuit dissection in squirrel monkeys. *Scientific Reports*, *8*(1), Article 6775. https://doi.org/10.1038/s41598-018-24362-7

Palminteri, S., Wyart, V., & Koechlin, E. (2017). The importance of falsification in computational cognitive modeling. *Trends in Cognitive Sciences*, *21*(6), 425–433. https://doi.org/10.1016/j.tics.2017.03.011

Pangalos, M. N., Schechter, L. E., & Hurko, O. (2007). Drug development for CNS disorders: Strategies for balancing risk and reducing attrition. *Nature Reviews. Drug Discovery*, *6*(7), 521–532. https://doi.org/10.1038/nrd2094

Paulus, M. P., Huys, Q. J., & Maia, T. V. (2016). A roadmap for the development of applied computational psychiatry. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *1*(5), 386–392. https://doi.org/10.1016/j.bpsc.2016.05.001

Pike, A. C., Lowther, M., & Robinson, O. J. (2021). The importance of common currency tasks in translational psychiatry. *Current Behavioral Neuroscience Reports*, *8*(1), 1–10. https://doi.org/10.1007/s40473-021-00225-w

Pisupati, S., Chartarifsky-Lynn, L., Khanal, A., & Churchland, A. K. (2021). Lapses in perceptual decisions reflect exploration. *eLife*, *10*, Article e55490. https://doi.org/10.7554/eLife.55490

Porter, J. N., Olsen, A. S., Gurnsey, K., Dugan, B. P., Jedema, H. P., & Bradberry, C. W. (2011). Chronic cocaine self-administration in rhesus monkeys: Impact on associative learning, cognitive control, and working memory. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *31*(13), 4926–4934. https://doi.org/10.1523/JNEUROSCI.5426-10.2011

Radulescu, A., & Niv, Y. (2019). State representation in mental illness. *Current Opinion in Neurobiology*, *55*, 160–166. https://doi.org/10.1016/j.conb.2019.03.011

Redish, A. D. (2004). Addiction as a computational process gone awry. *Science*, *306*(5703), 1944–1947. https://doi.org/10.1126/science.1102384

Remijnse, P. L., Nielen, M. M. A., van Balkom, A. J. L. M., Cath, D. C., van Oppen, P., Uylings, H. B. M., & Veltman, D. J. (2006). Reduced orbitofrontal-striatal activity on a reversal learning task in obsessive-compulsive disorder. *Archives of General Psychiatry*, *63*(11), 1225–1236. https://doi.org/10.1001/archpsyc.63.11.1225

Rescorla, R. A. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Current research and theory* (pp. 64–99). Appleton Century Crofts.

Samejima, K., Ueda, Y., Doya, K., & Kimura, M. (2005). Representation of action-specific reward values in the striatum. *Science*, *310*(5752), 1337–1340. https://doi.org/10.1126/science.1115270

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*(5306), 1593–1599. https://doi.org/10.1126/science.275.5306.1593

Soltani, A., & Izquierdo, A. (2019). Adaptive learning under expected and unexpected uncertainty. *Nature Reviews Neuroscience*, *20*(10), 635–644. https://doi.org/10.1038/s41583-019-0180-y

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. MIT Press.

Swainson, R., Rogers, R. D., Sahakian, B. J., Summers, B. A., Polkey, C. E., & Robbins, T. W. (2000). Probabilistic learning and reversal deficits in patients with Parkinson's disease or frontal or temporal lobe lesions: Possible adverse effects of dopaminergic medication. *Neuropsychologia*, *38*(5), 596–612. https://doi.org/10.1016/S0028-3932(99)00103-7

Takemoto, A., Miwa, M., Koba, R., Yamaguchi, C., Suzuki, H., & Nakamura, K. (2015). Individual variability in visual discrimination and reversal learning performance in common marmosets. *Neuroscience Research*, *93*, 136–143. https://doi.org/10.1016/j.neures.2014.10.001

Tallman, J. F. (1999). Neuropsychopharmacology at the new millennium: New industry directions. *Neuropsychopharmacology*, *20*(2), 99–105. https://doi.org/10.1016/S0893-133X(98)00104-3

Taswell, C. A., Costa, V. D., Murray, E. A., & Averbeck, B. B. (2018). Ventral striatum's role in learning from gains and losses. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(52), E12398–E12406. https://doi.org/10.1073/pnas.1809833115

Tsutsui, K. I., Grabenhorst, F., Kobayashi, S., & Schultz, W. (2016). A dynamic code for economic object valuation in prefrontal cortex neurons. *Nature Communications*, *7*(1), Article 12554. https://doi.org/10.1038/ncomms12554

Wilson, R. C., & Collins, A. G. (2019). Ten simple rules for the computational modeling of behavioral data. *eLife*, *8*, Article e49547. https://doi.org/10.7554/eLife.49547

Received April 28, 2021
Revision received June 22, 2021
Accepted July 10, 2021 ■

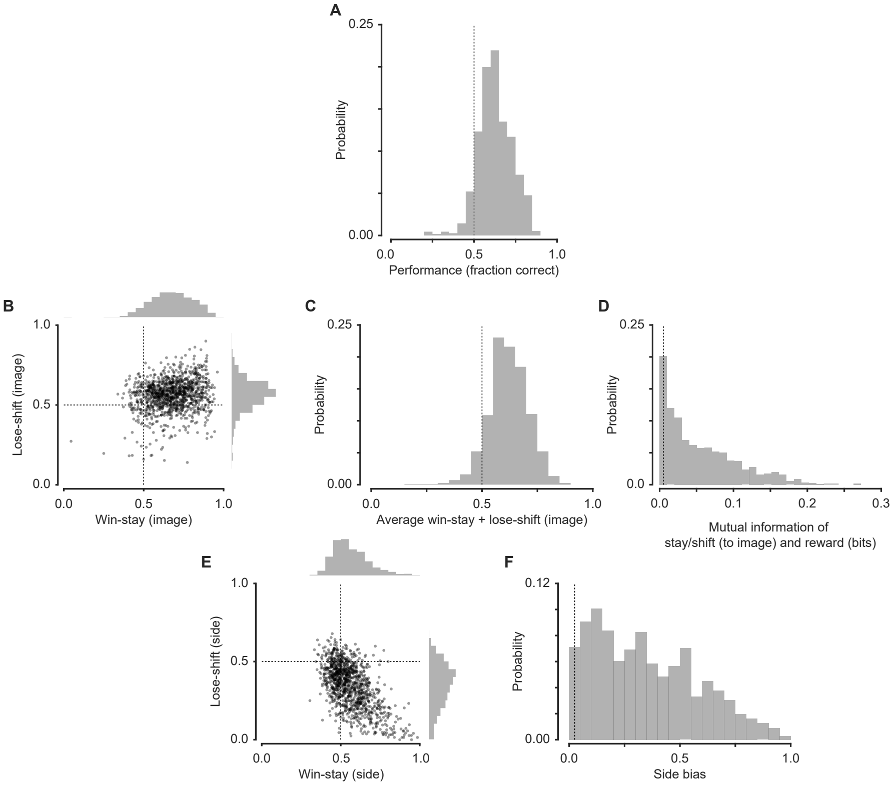


Figure S1: Individual session behavioral features demonstrate reward sensitivity and side bias. Behavioral features, plotted in a similar fashion to Figure 2. (A) Performance. (B) Image-based win-stay and lose-shift. Histograms above and to the right are marginal distributions. (C) Average win-stay + lose-shift. (D) Mutual information between stay/switch and reward on the previous trial. Dashed line is from simulated random behavior. (E) Side-based win-stay and lose-shift. Histograms above and to the right are marginal distributions. (F) Side bias.

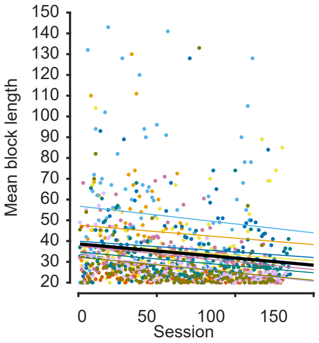


Figure S2: Mean block length decreases with training. The mean block length - another metric of performance - decreased significantly with training, mirroring the increase in fraction correct. Dots represent the mean block length in individual sessions. Black line shows the fixed effect and thin colored lines show individual monkey random effects. Colors denote individual monkeys and are consistent between figures.

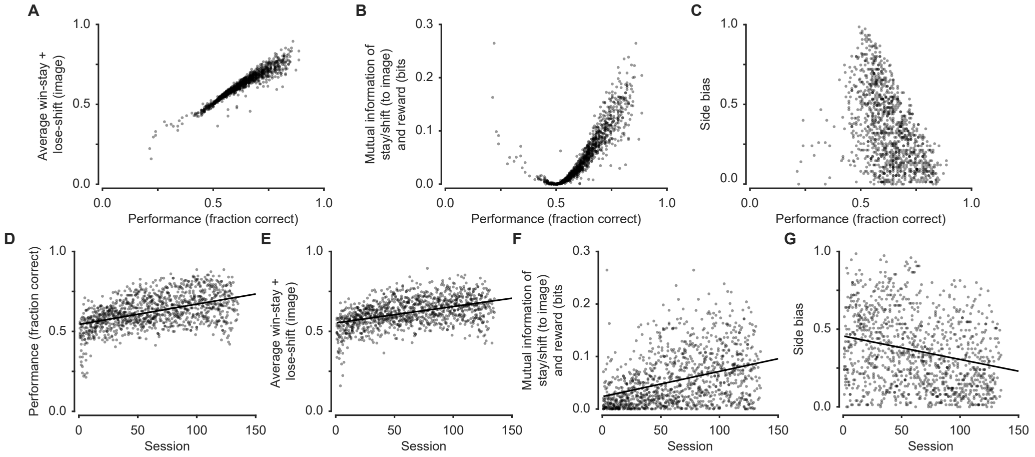


Figure S3: Individual session relationship between performance, reward sensitivity, choice bias, and training. Plotted in a similar fashion to Figure 3. (A) The average win-stay + lose-shift increased with increased performance. (B) The mutual information between stay/shift and reward increased with performance > 0.5. (C) Side bias was higher and more widely distributed when performance was closer to 0.5 and reduced in mean and variance when performance was better. (D) Performance improved with more sessions performed. (E) The average win-stay + lose-shift improved with training. (F) The mutual information between stay/switch and reward increased with training. (G) Side bias decreased with training.

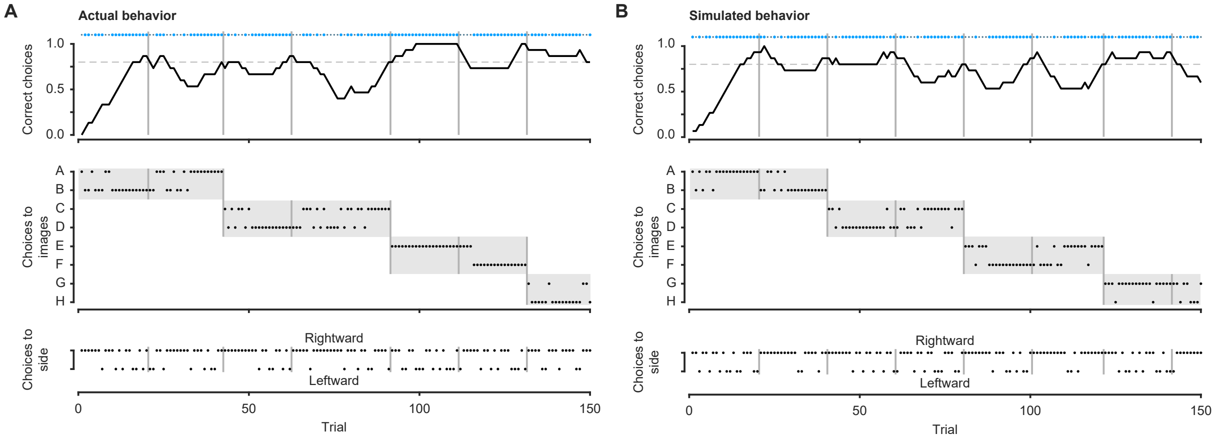


Figure S4: Raw behavior for actual and simulated sessions (A) Top panel: Correct choices as a function of trial. Large dots indicate a big reward choice and small dots indicate a small reward choice. The black line shows the fraction correct in the past 15 trials. The dashed line is the performance threshold (80% correct) used to trigger block transitions. Vertical grey lines indicate block transitions. Middle panel: Choices to images as a function of trial in the same format as 1C. Black dots indicate a choice to a respective image. Bottom: Choices to a side as a function of trial. Rightward (leftward) choices are indicated with a black dot on the top (bottom) of the figure. This session demonstrates a slight rightward side bias. (B) Behavior from the same session was fit to the reinforcement learning model to estimate parameters. These parameters were used to simulate an entirely new, synthetic data session.

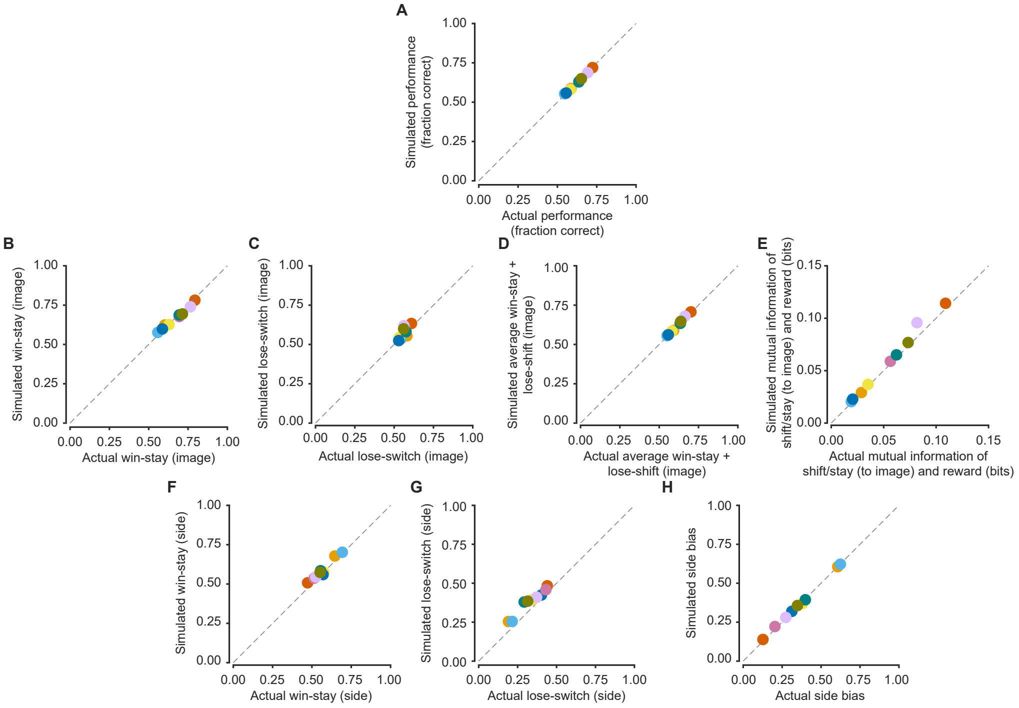


Figure S5: Comparison of behavioral features for actual and simulated behavior. To compare how well actual and simulated behavioral features match, we compute the mean difference between actual and simulated behavioral features [Mean (actual minus simulated) 95% CI] for each panel. (A) Average performance for each session. [−0.0015 – 0.0042] (B) Image-based win-stay [−0.0070 – 0.0130] (C) Image-based lose-shift [−0.0303 – 0.0015] (D) Image-based average win-stay + lose-shift [−0.0043 – 0.0064] (E) Mutual information of stay/shift and reward on the previous trial [−0.0082 – −0.0021] (F) Side-based win-stay [−0.0265 – −0.0062] (G) Side-based lose-shift [−0.0632 – −0.0367] (H) Side bias [−0.0097 – 0.0015]. Colors denote individual monkeys and are consistent between figures.

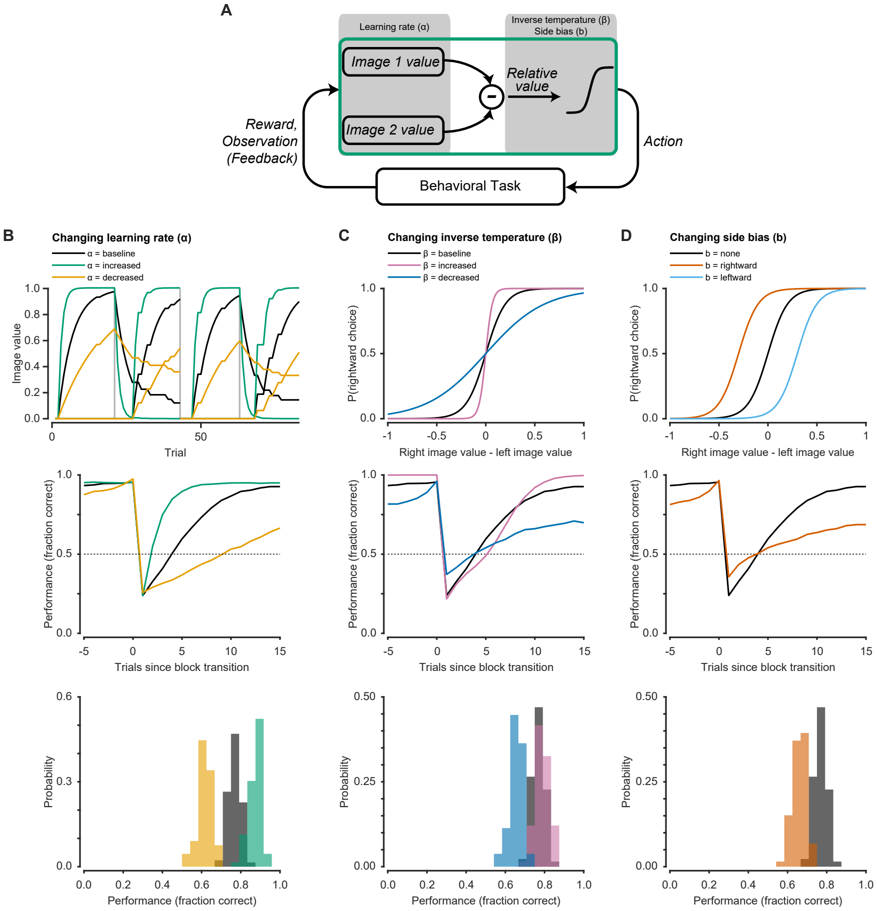


Figure S6: Model illustration and effects of varying parameters (A) Illustration of reinforcement learning model. Image values are updated by feedback via reward prediction errors (the discrepancy between predicted and actual rewards). This process is governed by the learning rate (α). The relative image value is mapped through a softmax function to produce a choice. This process is governed by the inverse temperature (β) and a side bias parameter. (B) Increasing the learning rate results in faster accumulation of reward value information. This results in faster block transitions and better overall performance. Decreasing the learning rate has the opposite effect. (C) Increasing the inverse temperature results in more deterministic choice behavior. Decreasing the inverse temperature makes choices more random. In this example, increasing the inverse temperature results in slower block transitions but more deterministic behavior after enough trials have elapsed, resulting in improved performance. Decreasing the inverse temperature has the opposite effect. Unlike the learning rate, the optimal inverse temperature is not at an extreme value but depends on the trials to criterion. Greater trials to criterion will favor a larger β. (D) Side bias results in increased choices of one particular side. Side bias is purely maladaptive and results in poorer overall performance.

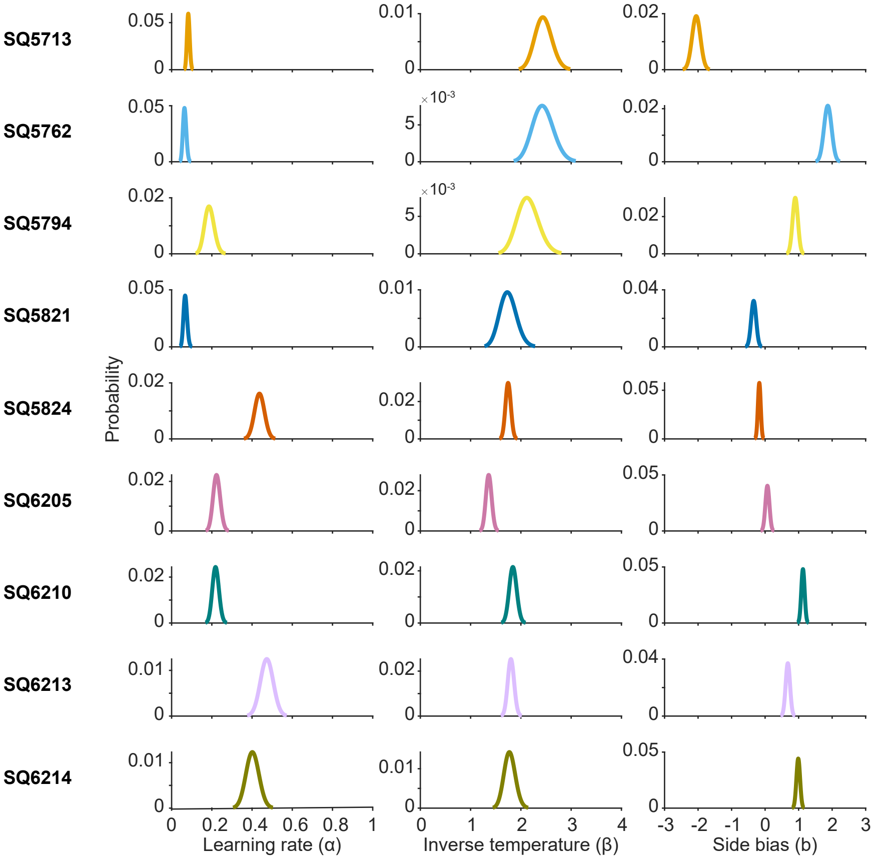


Figure S7: Parameter estimates for each monkey. Estimated learning rates, inverse temperatures, and side biases for all monkeys included in this study. These are monkey-level distributions, from which session-level parameters are drawn. Colors indicate the color used for that monkey throughout figures.

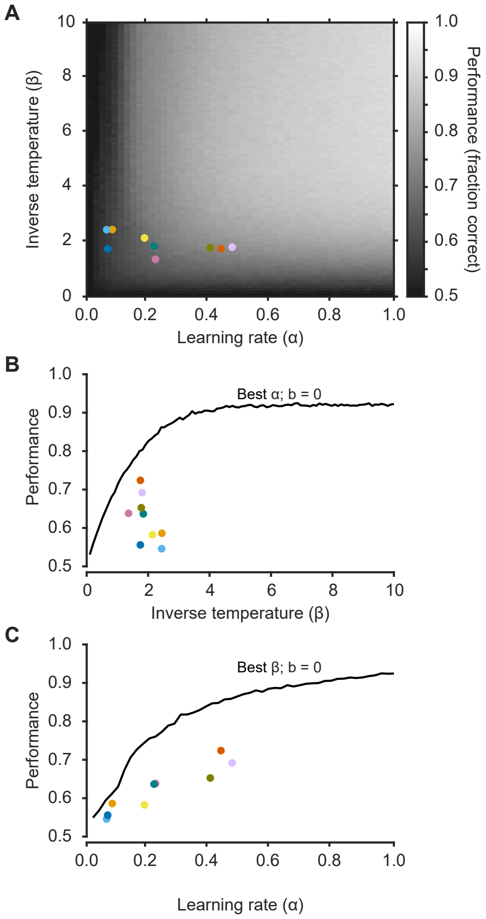


Figure S8: Performance as a function of learning rate (α) and inverse temperature (β). Each simulation was run for 66 sessions, each 2000 trials long, over 50 α values, 100 β values, and side bias fixed at 0. (A) Heatmap of performance for combinations of learning rates and inverse temperatures, with side bias fixed at 0. Performance is poor at low learning rates (regardless of the inverse temperature) and low inverse temperatures (regardless of the learning rate). In general, there is a large range of learning rates and inverse temperatures that permits adaptive behavior. Individual monkeys are shown with colored dots. Monkeys consistently maintain a suboptimal α/β combination. (B) Performance as a function of inverse temperature for the best learning rate and side bias $= 0$. Optimal performance is achieved at $\beta \gtrsim 4$. (C) Performance as a function of learning rate for the best inverse temperature and side bias $= 0$. Optimal performance is achieved at $\alpha = 1$. Colors denote individual monkeys and are consistent between figures.

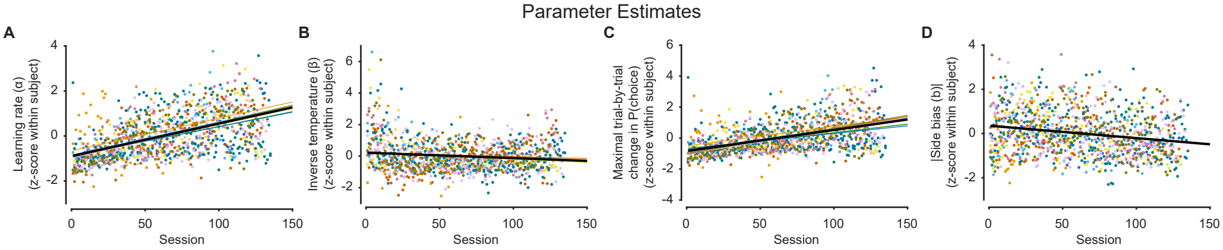


Figure S9: Within-subject normalization of parameters results in similar changes with training (A) Normalized learning rates improved with training (linear slope 1.44×10^{-2}, $t_{1091} = 17.33$, p < 0.0001). (B) Normalized inverse temperatures decreased with training (linear slope -3.46×10^{-3}, $t_{1091} = -4.05$, p < 0.0001). (C) Normalized maximal change in P(choice) increased with training (linear slope 1.36×10^{-2}, $t_{1091} = 11.13$, p < 0.0001). (D) Normalized absolute side bias decreased throughout training (linear slope -5.61×10^{-3}, $t_{1091} = -7.11$, p < 0.0001). Colors denote individual monkeys and are consistent between figures.

Monkey	Model									
	RW + Side bias (3 parameters)		RW + Side bias + Reversal mechanism (3 parameters)		RW + Side bias + Forget unchosen (4 parameters)		RW + Side bias + Forget unchosen + reward coded as [0.25 1] (4 parameters)		Side bias (1 parameter)	
	LH	BIC	LH	BIC	LH	BIC	LH	BIC	LH	BIC
SQ5713	3685	8363	3756	8505	3639	8601	3638	8598	4586	9503
SQ5762	6271	14198	6366	14389	6152	14513	6114	14438	7288	15129
SQ5794	9139	20186	9239	20385	8833	20211	8791	20127	10595	21826
SQ5821	11387	24796	11512	25046	11310	25315	11272	25240	12446	25566
SQ5824	9588	21135	9522	21003	9260	21132	9285	21181	13089	26831
SQ6205	10776	23475	10801	23525	10583	23728	10585	23732	12573	25786
SQ6210	9483	20988	9559	21140	9199	21093	9223	21141	11835	24344
SQ6213	9341	20675	9313	20619	8750	20157	8777	20212	12556	25777
SQ6214	8535	18925	8591	19037	8091	18655	8085	18645	11073	22764

Table S1: Model comparison. Comparison of negative log likelihood (LH) and Bayesian information criterion (BIC) values for the four models best fit to at least one monkey, and one noise model. The four best-fit models are variations of the Rescorla-Wagner (RW) model with a side bias. The noise model is a side bias only model. LH and BIC values are sums across all sessions for individual monkeys. Colors in the Monkey column indicate the color used for that monkey throughout figures. Gray highlights the best model (smallest BIC) for each monkey.