Article

Neuron

Metaplasticity as a Neural Substrate for Adaptive Learning and Choice under Uncertainty

Highlights

- Metaplasticity can adjust learning according to reward uncertainty
- Learning is adjusted without knowledge of the environment or explicit optimization
- Metaplasticity model predicts choice behavior more accurately than optimal models
- Changes in the activity of model neurons can be used to estimate uncertainty

Authors

Shiva Farashahi, Christopher H. Donahue, Peyman Khorsand, Hyojung Seo, Daeyeol Lee, Alireza Soltani

Correspondence

soltani@dartmouth.edu

In Brief

Farashahi et al. show that reward-dependent metaplasticity provides a robust mechanism for reward integration under uncertainty and for estimation of uncertainty. This synaptic mechanism allows reward feedback to adjust learning without any explicit optimization or knowledge of the task structure.

Farashahi et al., 2017, Neuron *94*, 401–414
April 19, 2017
http://dx.doi.org/10.1016/j.neuron.2017.03.044

CellPress

Neuron
Article

Metaplasticity as a Neural Substrate for Adaptive Learning and Choice under Uncertainty

Shiva Farashahi,[1] Christopher H. Donahue,[2,3] Peyman Khorsand,[1] Hyojung Seo,[4] Daeyeol Lee,[3,4,5,6] and Alireza Soltani[1,7,*]

[1]Department of Psychological and Brain Sciences, Dartmouth College, NH 03755, USA
[2]The Gladstone Institutes, San Francisco, CA 94158, USA
[3]Department of Neuroscience, Yale School of Medicine, New Haven, CT 06510, USA
[4]Department of Psychiatry, Yale School of Medicine, New Haven, CT 06511, USA
[5]Kavli Institute for Neuroscience, Yale School of Medicine, New Haven, CT 06510, USA
[6]Department of Psychology, Yale University, New Haven, CT 06520, USA
[7]Lead Contact
*Correspondence: soltani@dartmouth.edu
http://dx.doi.org/10.1016/j.neuron.2017.03.044

SUMMARY

Value-based decision making often involves integration of reward outcomes over time, but this becomes considerably more challenging if reward assignments on alternative options are probabilistic and non-stationary. Despite the existence of various models for optimally integrating reward under uncertainty, the underlying neural mechanisms are still unknown. Here we propose that reward-dependent metaplasticity (RDMP) can provide a plausible mechanism for both integration of reward under uncertainty and estimation of uncertainty itself. We show that a model based on RDMP can robustly perform the probabilistic reversal learning task via dynamic adjustment of learning based on reward feedback, while changes in its activity signal unexpected uncertainty. The model predicts time-dependent and choice-specific learning rates that strongly depend on reward history. Key predictions from this model were confirmed with behavioral data from non-human primates. Overall, our results suggest that metaplasticity can provide a neural substrate for adaptive learning and choice under uncertainty.

INTRODUCTION

Our knowledge of the world is continuously modified by the outcomes of a myriad of decisions we make over time. For this learning to be successful, the brain has to adjust the way it responds to and integrates each reward outcome, since stimulus-reward or action-reward contingencies can unpredictably change over time in natural environments. For example, in an environment where reward probability associated with alternative choices changes frequently (i.e., volatile environment), learning should be fast so that only the most recent outcomes are considered in the estimation of reward probabilities. By contrast, in a stable environment where reward probabilities do not change very often, learning should be slow to obtain a more accurate estimate of reward values. Indeed, it has been experimentally demonstrated that humans are able to adjust their learning or learning rates based on the reward uncertainty and volatility in a given environment (Behrens et al., 2007; Krugel et al., 2009; Nassar et al., 2010). However, neural mechanisms underlying these adjustments are relatively unknown.

Some reinforcement learning (RL) models assume that the proper learning rates in a given environment can be estimated using optimization processes. An example of such an optimization process is meta-learning, which maximizes the average of reward via modification of the learning rates based on a comparison between the medium-term and long-term running averages of the reward rate (Doya, 2002; Schweighofer and Doya, 2003; Soltani et al., 2006). However, it is unclear whether such an algorithm can determine the proper learning rates in dynamic tasks because the time constants for computing these reward averages also have to be adjusted according to the uncertainty and volatility of the environment. In contrast, model-based or Bayesian approaches try to learn the structure of the environment and its parameters in order to optimally integrate reward feedback (Behrens et al., 2007; Courville et al., 2006; Daw et al., 2005, 2006; Hampton et al., 2006; Mathys et al., 2011; Nassar et al., 2010, 2012; Payzan-LeNestour and Bossaerts, 2011; Yu and Dayan, 2005). For example, a Bayesian learner can assume different levels of uncertainty in the environment to estimate reward probability optimally (Behrens et al., 2007). Other models deal with reward uncertainty by detecting changes in the environment to adapt learning (Costa et al., 2015; Jang et al., 2015; McGuire et al., 2014; Nassar et al., 2010, 2012). Despite the success of these Bayesian models in capturing behavioral data and observation of neural correlates of variables in these models, it is unclear how the correct model of the environment can be learned or how different components of these models can be implemented in the brain.

We hypothesized that certain brain areas might be endowed with synaptic mechanisms that use reward history to determine the level of synaptic plasticity and, therefore, adjust learning in the absence of any explicit optimization, or before a model of

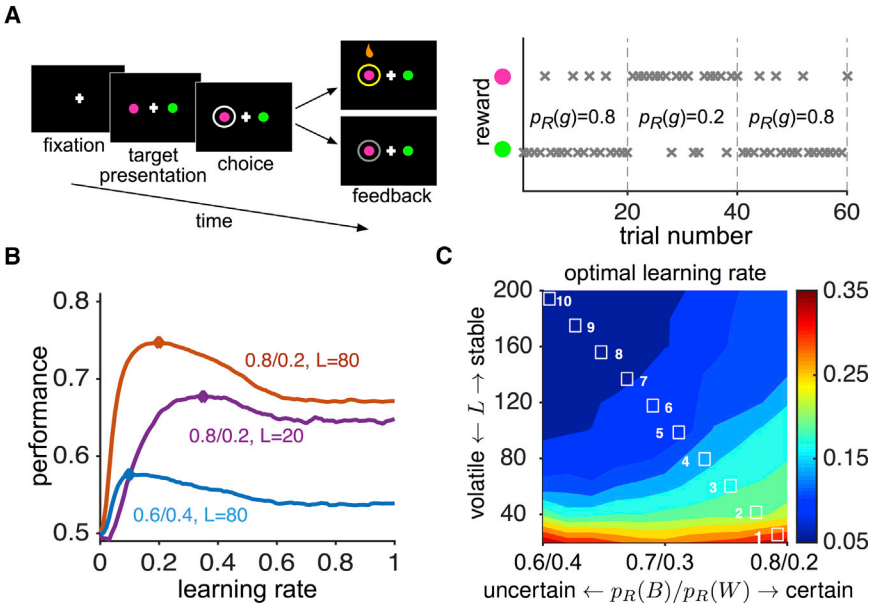

Figure 1. Probabilistic Reversal Learning Task and Reward Uncertainty

(A) Timeline of the PRL task and an example reward schedule. Subjects select between two options (e.g., red and green targets) and receive reward feedback on every trial. The reward is assigned probabilistically to one of the two targets while the better target changes between blocks of trials. In the shown example, probability of reward on the green target ($p_R(g)$) changes between 0.8 and 0.2 after every 20 trials. Each cross shows reward assignment on a given trial.

(B) Performance of the RL(1) model as a function of the learning rate in three environments with different levels of uncertainty or volatility. The diamond sign shows the optimal learning rate for each environment.

(C) The optimal learning rate for the RL(1) model in different environments quantified with reward probability on the better and worse options and the block length, L. The optimal learning rate was smaller for more stable, and to a lesser extent, for more uncertain environments. White squares indicate sample environments chosen for further tests.

the environment can be fully learned. Specifically, we assumed that such adjustment of learning can be implemented via synaptic metaplasticity. Metaplasticity refers to experimentally observed changes in the synaptic state that shape the direction, magnitude, and duration of future synaptic changes without any observable change in the efficacy of synaptic transmission (Abraham, 2008; Abraham and Bear, 1996; Hulme et al., 2013; Müller-Dahlhaus and Ziemann, 2015). Here, we present a biophysically plausible model based on reward-dependent metaplasticity (RDMP) that can adjust learning according to reward uncertainty. In our model, metaplastic synapses have multiple levels of stability (meta-states) associated with two levels of synaptic efficacy and undergo reward-dependent changes. To test our model, we simulated choice behavior during a dynamic learning and decision-making task known as probabilistic reversal learning (PRL), which has been extensively used to study adaptive learning and decision making in health and disease (Cools et al., 2001; Costa et al., 2015; Rudebeck et al., 2013; Rygula et al., 2010; Swainson et al., 2000).

In this study, we demonstrate that based on reward feedback, RDMP allows synapses to occupy states with different levels of stability and thus can adjust learning according to the uncertainty and volatility of the environment. We show that a model endowed with RDMP not only can perform the PRL task over a wide range of model and task parameters, but also provides specific predictions that differ from those of heuristic RL models, and Bayesian models that require explicit knowledge of the task structure. Using an extensive set of behavioral data from a recent study in non-human primates during the PRL task (Donahue and Lee, 2015), we provide experimental evidence for some of the model's predictions and thus, the contribution of metaplasticity to behavior. Finally, we show that changes in the activity of reward-encoding neurons with metaplastic synapses can be used to estimate reward volatility for which a neural correlate has been observed (Behrens et al., 2007). Overall, our results suggest that reward-dependent metaplasticity can provide a robust neural substrate for adaptive learning and choice under uncertainty, and moreover, enables computations of high-level signals such as volatility.

RESULTS

Learning and Choice Under Reward Uncertainty

As a platform to study learning and choice under reward uncertainty, we focused on behavior during a PRL task. In this task, the subject selects between two alternative options (e.g., colored targets), which deliver reward probabilistically (Figure 1A). The probabilities of reward on the green and red options are complementary, for example 0.8 on the green and 0.2 on the red target. However, unknown to the subject, these probabilities switch after a certain number of trials referred to as the block length, L. The combination of reward probability on the better (more rewarding) and worse (less rewarding) options, $p_R(B)$ and $p_R(W)$, and block length defines an environment in this task (e.g., 0.8/0.2 schedule with L = 80). Performing this task requires selection of the better option within a given block of trials, which is complicated due to two factors: (1) the probabilistic nature of reward assignment or expected "uncertainty"; and (2) switches in reward probabilities between blocks of trials (reversals), resulting in unexpected uncertainty, also referred to as "volatility."

To detect the better option within each block of trials in the PRL task, the subject has to continuously update the estimates for reward value of the two options based on reward feedback. However, for models with constant learning rates, such as a simple RL model with one learning parameter (RL(1)), the optimal learning rates depend on the levels of uncertainty and volatility in the environment (Figures 1B and 1C). In a relatively stable environment with reward probabilities far from 0.5 (0.8/0.2 schedule with L = 80), a moderate value of the learning rate produces optimal performance (Figure 1B). However, in a more volatile

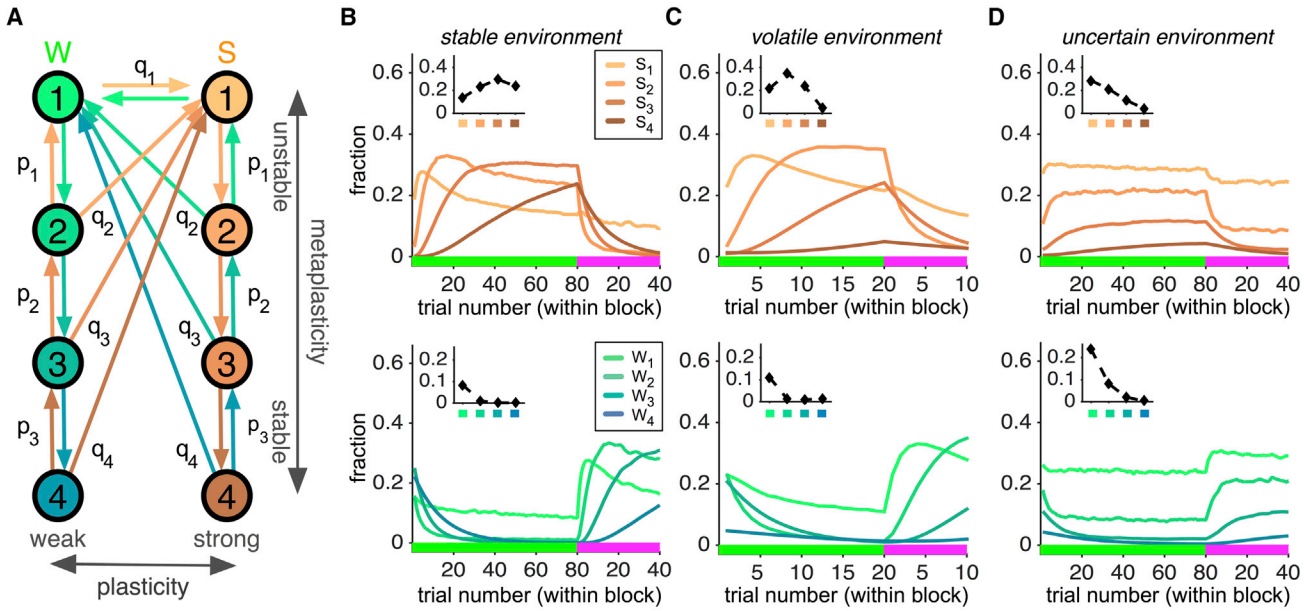

Figure 2. The RDMP Model and Its Response to Three Environments with Different Levels of Uncertainty or Volatility

(A) The schematic of metaplastic synapses. Metaplastic synapses have multiple meta-states associated with each of the two levels of synaptic efficacy: weak (W) and strong (S). Potentiation and depression events result in stochastic transitions between meta-states with different levels of stability and are indicated by arrows (in gold and cyan for potentiation and depression events, respectively) and quantified by different transition probabilities ($q_1 > q_2 > q_3 > q_4$ and $p_1 > p_2 > p_3$). We also refer to more unstable and stable meta-states as "shallower" and "deeper" meta-states, respectively.

(B) For synapses associated with the green target, the average (over many blocks) fractions of synapses in different strong (top) and weak (bottom) meta-states are plotted over time in the stable environment (0.8/0.2 schedule with $L = 80$). The x axis color indicates the better option within a given block and the inset shows the steady state of the fraction of synapses in each of four meta-states (computed by averaging over the last 2 trials within each block).

(C and D) The same as (B) but the results for the volatile (0.8/0.2 schedule with $L = 20$) (C) and uncertain environments (0.6/0.4 schedule with $L = 80$) (D) are shown.

environment (0.8/0.2 schedule with $L = 20$), a higher learning rate is more desirable. Finally, in an environment with greater uncertainty (0.6/0.4 schedule with $L = 80$), a lower learning rate is required to obtain a better estimate of reward value. Overall, the optimal learning rate of RL(1) increases as the environment becomes more volatile and slightly decreases when the environment becomes more uncertain (Figure 1C).

These results demonstrate an inherent tradeoff between adaptability (i.e., fast response to changes in the environment) and precision (i.e., correct estimation of reward probabilities) during learning and choice under uncertainty. Tackling this tradeoff requires the brain to adjust the learning rate either across time or across environments since the optimal learning rate could vary substantially depending on the levels of uncertainty and volatility present.

RDMP as a Neural Mechanism for Adaptive Learning Under Reward Uncertainty

Here, we suggest that RDMP can provide a plausible mechanism for learning and choice under uncertainty and simulate the behavior of an example model based on RDMP during the PRL task. We compare the behavior of this model with three sets of models that rely on different mechanisms to deal with uncertainty and volatility in the PRL task (see STAR Methods). In our model, neurons encoding the reward value of different options (value-encoding neurons) receive their inputs via metaplastic synapses that undergo a stochastic RDMP learning rule. These metaplastic synapses have multiple meta-states with different levels of stability associated with two levels of synaptic efficacy: weak and strong. The output of value-encoding neurons associated with a given option reflects the overall synaptic efficacy of metaplastic synapses onto them. Because there are two levels of synaptic efficacy, the overall synaptic efficacy for each set of metaplastic synapses can be quantified as the fraction of synapses in strong meta-states, which we refer to as the "synaptic strength." Importantly, the RDMP learning rule enables the synaptic strengths to estimate the probability of reward for alternative options, whereas the model selects between the two options stochastically based on a probability given by the difference in synaptic strengths for the two options (see STAR Methods).

Metaplastic synapses can change their states stochastically depending on the choice and reward outcome at the end of each trial. We assumed that synapses associated with the chosen option are potentiated on rewarded trials and depressed on unrewarded trials (Figure 2A; see STAR Methods). Because only one of the two options is assigned with reward on each trial of the PRL task, we also assumed that synapses associated with the unchosen option are depressed on rewarded trials and potentiated on unrewarded trials. Due to the stochastic nature of synaptic transitions, potentiation or depression events may or may not change synaptic efficacy of a particular synapse, but at the population level, result in a well-defined learning rule (Equations 3 and 4 in STAR Methods).

We simulated the behavior of the model in a few environments with different levels of uncertainty and volatility, using the same set of parameters. Toward the end of each block in the stable environment, synapses associated with the better option increasingly occupied more stable strong meta-states, while a small fraction of them occupied the most unstable weak meta-states (W1; Figure 2B). The volatile environment, on the other hand, made these synapses occupy mainly unstable strong meta-states or the most unstable weak meta-states (Figure 2C). This happened because, in the volatile environment, there was not enough time for synapses undergoing potentiation to occupy stable strong meta-states, whereas this was possible in the stable environment. In addition to the block length, the fractions of synapses in different meta-states were also influenced by reward probabilities (hence uncertainty). As the reward probabilities became closer to 0.5, thus increasing uncertainty, more synapses occupied more unstable weak meta-states and the fractions of synapses in strong meta-states monotonically decreased for more stable meta-states (Figure 2D). This happened because with more variable reward assignment, synapses associated with the better option were less likely to transition to stable (deep) meta-states. Overall, these results indicate that metaplastic synapses can adjust to reward statistics in the environment, in terms of both the volatility and uncertainty.

Adjustment of Learning Over Time

The fractions of synapses in different meta-states show how metaplastic synapses adjust to reward statistics in a given environment. Because different meta-states have different transition probabilities, those fractions also determine the speed of learning at a given point in time. To illustrate these, we calculated the ''effective'' learning rates as a function of the trial number after a reversal for the two possible outcomes of reward assignment. For any point during a block, the effective learning rates provide a single set of learning rates by considering the total change in the efficacy over all meta-states (see STAR Methods). By definition, the product of the effective learning rate and the fraction of synapses in weak (strong) meta-states are equal to the increase (decrease) in synaptic strength.

We found that the effective learning rates changed over time and depended on whether the reward was assigned to the better or worse option. For synapses associated with the better option, the effective learning rate on trials when the worse option was assigned with reward, $K_{B-}(t)$, monotonically decreased over time after a reversal (solid cyan curve in Figure 3A). At the same time, however, the effective learning rate on trials when the better option was assigned with reward, $K_{B+}(t)$, monotonically increased. The amount of changes in the effective learning rates depended on the uncertainty and volatility such that these changes were larger for more certain and stable environments (see below).

At the beginning of each reversal in the stable environment, synapses associated with the better option in the new block were mainly in stable weak meta-states or unstable strong meta-states since these synapses were associated with the worse option in the previous block (Figure 2B). On trials when reward was assigned to the better option, synapses in the stable weak meta-states slowly transition to strong meta-states, resulting in a small effective learning rate at the beginning of each block (solid gold curve, Figure 3A). In contrast, on trials when reward was assigned to the worse option, synapses in the unstable strong meta-states quickly transition to weak meta-states resulting in a large value for the effective learning rate on those trials. Both of these effective learning rates change over time as the distribution of synapses in different meta-states adjusts to the recent reward statistics (Figures 2B–2D).

The change in the effective learning rates over time as well as the difference between the two learning rates were sensitive to reward uncertainty and volatility in the environment. Specifically, the difference between $K_{B+}(t)$ and $K_{B-}(t)$ was more pronounced in a more certain environment than in an uncertain one (compare solid and dashed curves in Figure 3A). Moreover, $K_{B+}(t)$ rose to a higher value while $K_{B-}(t)$ fell to a lower value in a stable than in a volatile environment (compare solid curves in Figure 3A and its inset). The time-dependent adjustment to reward statistics was not specific to the example environments and was observed over a large set of environments with different levels of uncertainty and volatility. At the beginning of each block, $K_{B+}(t)$ was smaller than $K_{B-}(t)$ and this difference was stronger (more negative) for more certain or stable environments (Figure 3B). Over time, $K_{B+}(t)$ increased while $K_{B-}(t)$ decreased such that $(K_{B+} - K_{B-})$ becomes positive later in the block (Figures 3C and 3D). The difference between the steady state of the two learning rates increased as uncertainty and/or volatility decreased.

Although our model based on RDMP suggests that learning rates depend on whether the reward outcome supports the better or worse choice alternative, these rates are often estimated in empirical studies based on the reward outcomes independently of choice alternatives. To estimate such learning rates in our model, we computed the effective learning rates on rewarded and unrewarded trials, $K_{rew}(t)$ and $K_{unr}(t)$, by averaging the effective learning rates based on the reward outcome on a given trial (Equation 8 in STAR Methods). We found that $K_{rew}(t)$ was smaller than $K_{unr}(t)$ at the beginning of each block (Figure S1). However, as the model spent more time in a block, $K_{rew}(t)$ increased, whereas $K_{unr}(t)$ decreased such that $K_{rew}(t)$ became larger than $K_{unr}(t)$ later in a block. These changes resulted in an overall larger learning rate for rewarded than unrewarded trials, as observed in previous experiments (Donahue and Lee, 2015; Frank et al., 2007, 2009; Niv et al., 2012). In addition to changes in the learning rates over time (i.e., trial to trial), our model also predicts that the difference between the overall learning rates on rewarded and unrewarded trials should decrease with the uncertainty in the environment (Figure S1F). This prediction can be tested in future experiments.

Overall, these results show that metaplastic synapses adjust to reward statistics in the environment. This gives rise to time-dependent learning rates that are different for synapses associated with the two alternative options (i.e., learning rates are choice specific), and on rewarded and unrewarded trials. For simplicity, here we used a specific implementation of RDMP (Equation 2 in STAR Methods), which guarantees that at the steady state, the effective learning rate for reward assignment on the better option is larger than the one on the worse option. Nevertheless, we found that most RDMP models with different formulations of transition probabilities exhibit such behavior, as long as there is an order

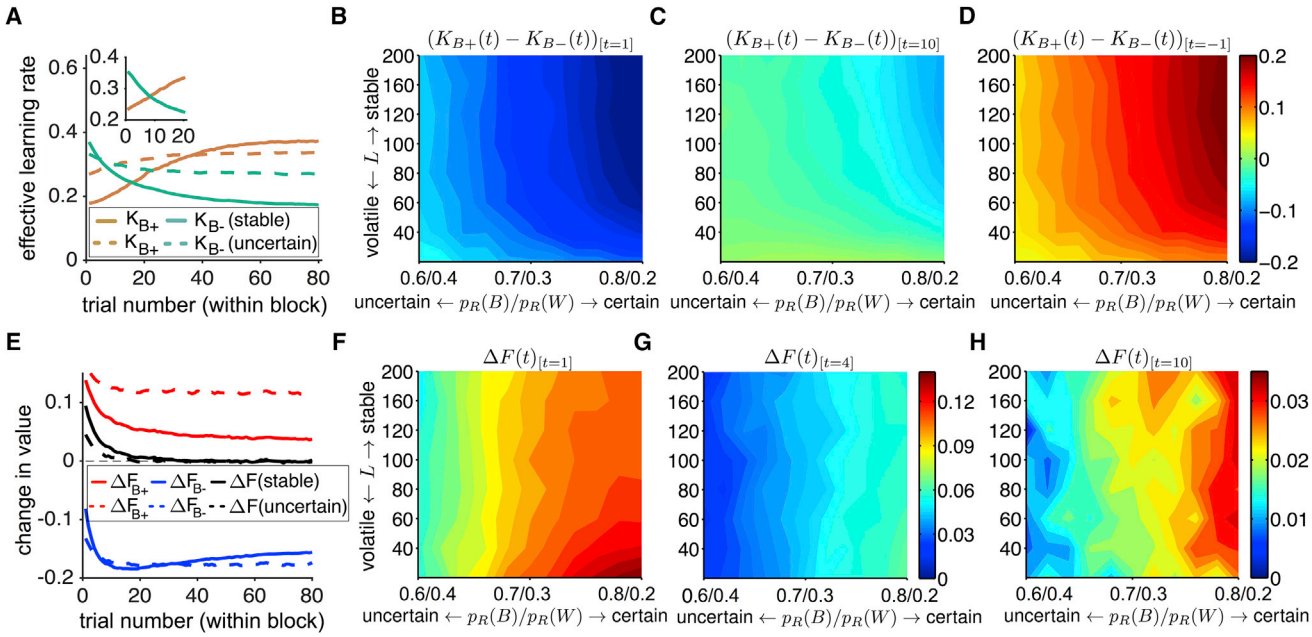

Figure 3. The RDMP Model Adjusts Learning Over Time According to Reward Uncertainty and Volatility

(A) The time course of the effective learning rate for when the reward was assigned to the better (K_{B+}) or worse (K_{B-}) option during a given block in the stable (0.8/0.2 schedule with L = 80) and uncertain (0.6/0.4 schedule with L = 80) environments. The inset shows the results for the volatile environment (0.8/0.2 reward schedule with L = 20).

(B–D) The difference between the effective learning rates at three time points after a reversal in different environments. Overall, K_{B+} increased while K_{B-} decreased and their difference was larger for more certain and/or stable environments.

(E) Changes in model's response to reward feedback over time. Plotted are the changes in the synaptic strength in response to reward assignment on the better (ΔF_{B+}) or worse option (ΔF_{B-}), as well as the overall change in the synaptic strength (ΔF) as a function of the trial number after a reversal in the stable and uncertain environments.

(F–H) The overall change in the synaptic strength at three time points after a reversal in different environments. The model's response to reward feedback was stronger for more certain and/or volatile environments right after reversals and this difference slowly decreased over time.

in the transitions between meta-states resulting in shallow and deep meta-states (data not shown).

Adjustments of Model's Response to Reward Uncertainty and Volatility

We next examined how the model endowed with metaplasticity can adjust its response according to the uncertainty and volatility in the environment. To do so, we computed changes in the model's response to reward feedback over time. Because choice behavior is determined by the synaptic strengths (Equation 1 in STAR Methods), we first computed changes in the synaptic strengths due to the two types of reward feedback at different time points within a block of the PRL task.

We found that the change in the synaptic strength when reward was assigned to the better option, $\Delta F_{B+}(t)$, was large immediately after a reversal, but then it slowly decreased over time (red curves in Figure 3E). This happens because immediately after a reversal, a large fraction of synapses associated with the currently better (previously worse) option is in weak meta-states, and the transition of these synapses due to a potentiation event results in a large $\Delta F_{B+}(t)$. The $\Delta F_{B+}(t)$ gradually decreases as fewer synapses remain in weak meta-states. On the other hand, the change in the synaptic strength when reward was assigned to the worse option, $\Delta F_{B-}(t)$, became stronger (more negative) over the span of about ten trials after a reversal, and later gradually became weaker (blue curves in Figure 3E). Importantly, the starting points of $\Delta F_{B+}(t)$ and $\Delta F_{B-}(t)$ were farther from zero but changed less over time in the uncertain compared to stable environment (compare dashed and solid curves in Figure 3E). In contrast, we observed larger changes in response to reward feedback over trials within each block of the volatile environment (data not shown).

To measure the model's overall response to both types of reward feedback, we also computed a weighted average of the values of $\Delta F_{B+}(t)$ and $\Delta F_{B-}(t)$ based on the reward probability in a given block (see STAR Methods). In the stable environment, $\Delta F(t)$ slowly decreased to zero after each reversal as the model reached the steady state within each block (solid black curve in Figure 3E). In the uncertain environment, however, $\Delta F(t)$ was initially lower and decreased to zero more slowly (dashed black curve in Figure 3E). These results demonstrate the overall ability of our model to adjust its response based on reward uncertainty in the environment.

To further examine adjustments due to metaplasticity, we simulated our model in various environments with different levels of uncertainty and volatility using one set of parameters. Immediately after each reversal, the model showed the greatest overall response to reward feedback; this response was larger for

more certain and/or volatile environments (Figure 3F). As the overall response to reward feedback gradually approached zero, it still remained sensitive to the level of uncertainty in the environment (Figures 3G and 3H). Our model's response to reward feedback was different from that of the RL models with constant learning rates. For example, in the RL(1) model, the change in the reward value due to reward feedback was similar in the stable and volatile environments (Figure S2A). Thus, unlike our model, the RL(1) model with a constant learning rate cannot adjust to the volatility in the environment.

Considering the observed adjustments in our model, we then compared our model's overall response to both types of reward feedback using one set of parameters to that of the RL(1) model with the optimal learning rate in each environment (Figures S2B–S2D). The dependency on the uncertainty and volatility was qualitatively similar between the two models (compare Figures 3F–3H with Figures S2B–S2D). These results show that metaplasticity enables our model to adjust its behavior to the uncertainty and volatility consistently with the RL model with the optimal learning rate; that is, to increase response to reward feedback in more certain or volatile environments. Our model's behavior, however, is not optimal and therefore shows deviations from what is prescribed by the optimal RL(1) model.

To achieve optimality, the RL(1) model prescribes smaller learning rates for more stable, and to a lesser extent, for more uncertain environments (Figure 1C). Such adjustment of the learning rate across environments is very different from how our model adjusts learning. First, our model naturally adopts two different time-dependent learning rates for reward assignments on the better and worse options. Second, the adjustment of these learning rates over time is qualitatively similar for more uncertain and more volatile environments (Figures 3B–3D), whereas the RL(1) model prescribes that the optimal learning rate should increase with larger volatility but slightly decrease with higher uncertainty (Figure 1C). Nevertheless, the opposite adjustment for uncertainty only weakly affects the performance in our model. This is because smaller differences between the fractions of synapses in the weak and strong meta-states in more uncertain environments cause smaller responses to reward feedback than in certain environments, irrespectively of the learning rates (Figure 3F). Similar behavior occurs in the RL models due to a smaller reward prediction error in uncertain compared to certain environments.

Our proposed RDMP model relies on an ordered architecture for transitions such that there are "shallow" and "deep" meta-states in the model. This architecture predicts that the model should be sensitive to the exact sequence of reward assignment. We found that after a sequence of consecutive reward assignments on the better option, the model responded very differently to another reward on the better option (congruent trial) versus reward on the worse option (incongruent trial), depending on the volatility of the environment (Data S1 and Figure S3). Importantly, these responses were qualitatively different from those of the RL and hierarchical Bayesian models.

To summarize, we show that reward-dependent metaplasticity offers a plausible solution for the integration of reward in environments with different levels of uncertainty and/or volatility. The RDMP model predicts that the learning rates change over time and are different depending on whether the reward is assigned to the better or worse option. Moreover, the model predicts a specific pattern of response after congruent and incongruent sequences of reward assignment, which is qualitatively different from those of alternative models.

Experimental Evidence

We next tested the predictions of our model by analyzing experimental data from a modified version of the PRL task in which monkeys selected between two color targets, which provided reward with different probabilities and magnitudes (Donahue and Lee, 2015). In this task, the reward was probabilistically assigned to the two targets similarly to the original PRL task, while the magnitude of reward was selected randomly from a set of four values (1, 2, 4, and 8 drops of juice) in order to encourage the animal to more equally select between the two targets (Donahue and Lee, 2015). Nevertheless, in order to successfully perform this task, the animal had to learn the probability of reward within each block by integrating reward feedback since the reward magnitudes were not predictive of reward assignment.

The first prediction of our model was that the learning rates change over time and differ depending on whether reward was assigned to the better or worse option. To test these predictions, we fit the choice behavior of monkeys with eight models (see STAR Methods). These include two Bayesian models with different mechanisms for solving the PRL task (hierarchical and change-detection Bayesian); two types of RL models, RL(1) and RL(2), with constant or time-dependent learning rates; the RDMP model with three meta-states; and a simplified version of the RDMP model. We used the simplified RDMP (sRDMP) to circumvent degeneracy in the solution (i.e., lack of unique solution) for the RDMP model, because each value of the synaptic strength could potentially correspond to many different distributions of synapses in different weak and strong meta-states. Moreover, the simplified RDMP is based on the critical prediction of the general metaplasticity model (different time-dependent effective learning rates for reward on the better and worse options) and thus can be used to detect behavioral contributions of RDMP independently of its specific implementation. To compare different models, we applied a 5-fold cross-validation, using the fit from 80% of the data from a given environment (95 sessions, about 53,000 trials in total) to predict the choice on the remaining 20% (23 sessions, about 13,000 trials in total). Crucially, the large amount of behavioral data allowed us to accurately test different models.

Overall, the RDMP and sRDMP models predicted choice behavior better than any of the competing models in both volatile and stable environments (Figure 4A). In contrast, the hierarchical Bayesian and the change-detection Bayesian models with optimal performance in the PRL task provided the worst fit to our experimental data. This illustrates that subjects did not perform optimally in the task. Moreover, the RL models with time-dependent learning rates predicted choice behavior better than the RL models with constant learning rates, indicating that learning rates adjusted over time. We also examined the average goodness-of-fit over trials within a block. This analysis revealed that our models predicted the choice behavior better than competing models, especially immediately after reversals

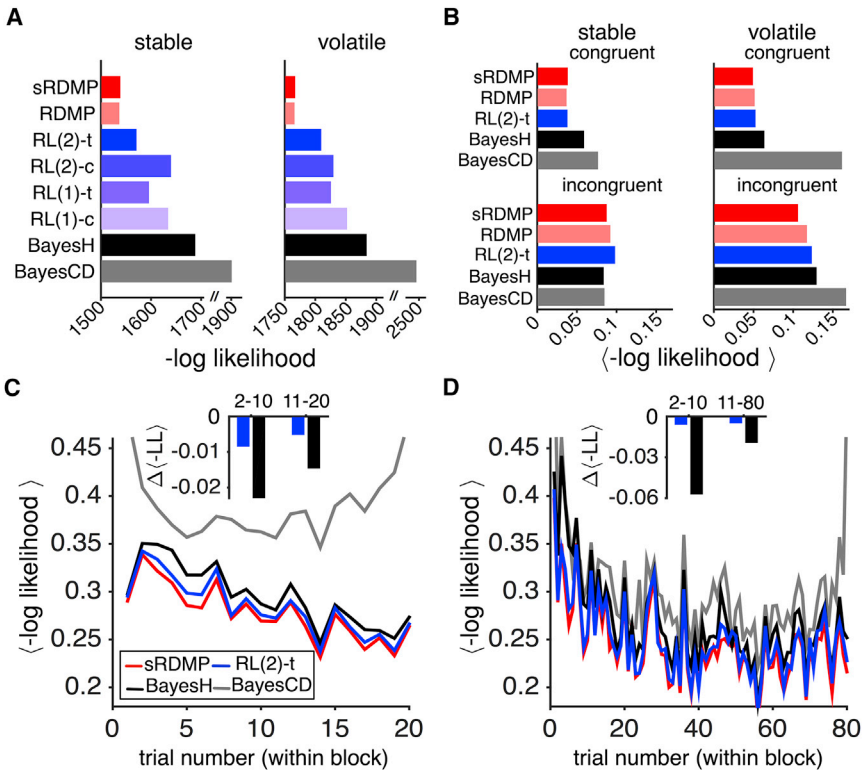

Figure 4. Model Comparison

(A) Comparison of the goodness of fit for monkeys' choice behavior during the modified PRL task using eight different models (BayesH: hierarchical Bayesian; BayesCD: change-detection Bayesian; RL-c and RL-t refer to RL models with constant and time-dependent learning rates). Plotted is the average negative log likelihood (-LL) over all cross-validation instances (using test trials) separately for data in the stable and volatile environments. Overall, the RDMP and sRDMP models provide the best fit in both environments, whereas the Bayesian models provide the worst fit.

(B) Goodness of fits for congruent and incongruent trials. For clarity, only the results for the best RL model are shown.

(C and D) Goodness of fits across time for different models during the volatile (C) and stable (D) environments. Plotted is the goodness of fit across time measured as the average -LL per trial, on a given trial within a block (based on cross-validation test trials). The blue (black) bars in the inset show the difference between the average -LL of sRDMP and RL(2)-t (respectively, hierarchical Bayesian) in early (trial 2-10) and late (trial 11-20, or 11-80) trials after a reversal. Overall, the sRDMP and RDMP (not shown to avoid clutter) models provide the best fit especially right after reversals.

(Figures 4C and 4D). This is important, because our model switches its behavior after reversals more slowly than other models (compare Figures 3E–3H to Figure S2) but this suboptimal response captures behavioral data.

The fit based on the sRDMP model also revealed significant changes in the learning rates over time, as predicted by our model. Namely, the average estimate of learning rate for trials when reward was assigned to the better target (K_{B+}) increased over time within a block, whereas the learning rate on trials when reward was assigned to the worse target (K_{B-}) decreased (Figures 5A and 5D). In contrast, the estimated learning rates using the RL models with time-dependent learning rates showed small changes over time (Figure S4). Interestingly, the fit based on the RDMP model revealed similar patterns for the estimated transition probabilities in the stable and volatile environments (Figures 5B and 5E). This indicates that similar sets of parameters could have been used to perform the task in both environments. Moreover, the striking similarity between the estimated learning rates based on sRDMP and the effective learning rates based on the estimated transition probabilities in the RDMP model (compare Figures 5A and 5D with Figures 5C and 5F) shows the feasibility of our approach in capturing the behavioral signature of metaplasticity without using a specific implementation of it. Together, these results strongly support our main predictions that the effective learning rates change over time and are different when reward was assigned to the better and worse options.

We also examined the second prediction of the RDMP model regarding congruent and incongruent trials, which distinguishes our model from the hierarchical Bayesian and RL(2) models, respectively (Figure S3). More specifically, we computed the average goodness of fit on trials following congruent and incongruent sequences of reward assignment in each environment. The RDMP, sRDMP, and RL(2) models predicted the monkeys' choices on trials following congruent sequences of reward assignment more precisely than the Bayesian models in both environments (Figure 4B). On incongruent trials, metaplasticity and Bayesian models provided better fits than the RL(2) model in the stable environment. In the volatile environment, however, the Bayesian models predicted monkeys' choices on incongruent trials less precisely than the RL(2) model. Overall, the monkeys' choice behavior was captured better with our models based on RDMP than the three competing models on both sequences of reward assignment.

Model's Robustness

To test the robustness of our model, we first simulated its behavior in ten separate environments that required very different optimal learning rates. These environments are defined with a given $p_R(B)/p_R(W)$ and L and are labeled in Figure 1C with white squares. Our model's simulation used one set of parameters and the resulting performance was compared with that of the Bayesian models and that of RL models using the optimal learning rates chosen separately for each environment (Figure 6A). We found that even using a single set of parameters, the RDMP model was able to perform only slightly below the hierarchical Bayesian model and RL models that used optimal learning rates in each environment. However, this does not

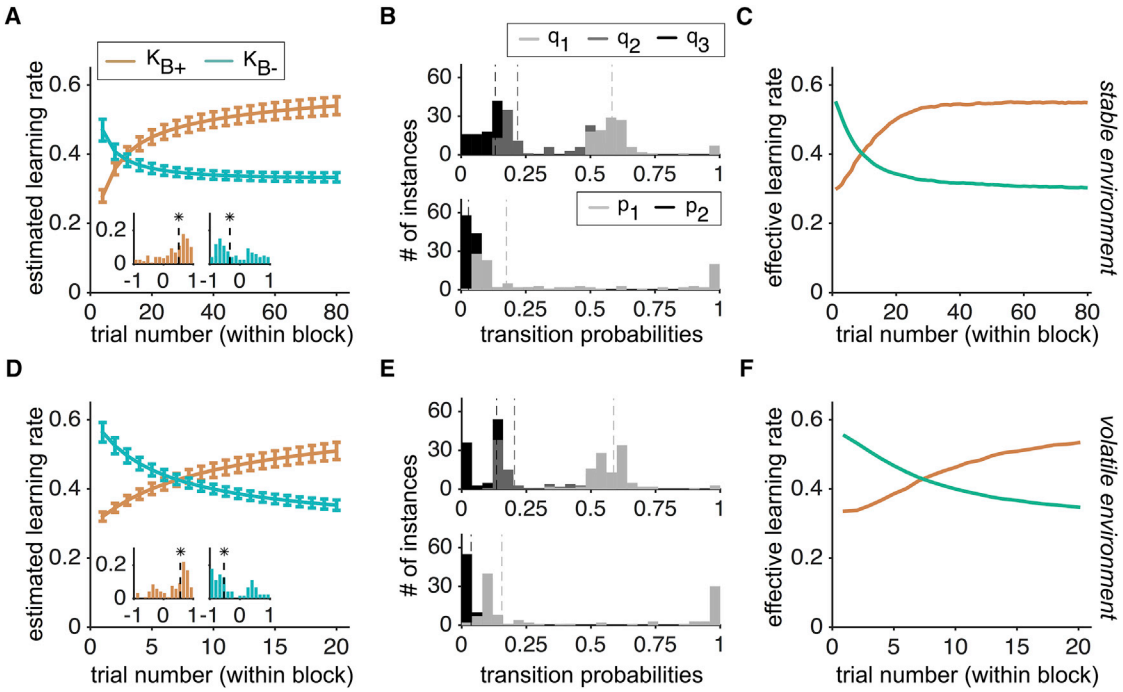

Figure 5. Experimental Evidence for Metaplasticity Revealed by Time-Dependent, Choice-Specific Learning Rates

(A) Plotted are the average estimated learning rates over time on trials when the reward was assigned to the better and worse options. These estimates are obtained using the session-by-session fit of monkeys' choice behavior with the sRDMP model in the stable environment. The error bars indicate SEM. The insets show the distributions of the difference between the steady state and initial values of the learning rates across all sessions (separately for each learning rate), and stars show whether the median (black dashed line) of each distribution is significantly different from zero ($p < 0.05$).

(B) The distribution of five transition probabilities estimated from fitting the choice behavior using the RDMP model with three meta-states ($m = 3$). Dashed lines show the median. The bimodal distribution for p_1 values is an indicative of degeneracy in the solution for the RDMP model.

(C) The effective learning rates in the RDMP model based on the median of estimated transition probabilities shown in (B).

(D–F) The same as (A)–(C) but for behavior in the volatile environment. Estimated transition probabilities and the effective learning rates showed similar patterns in the two environments.

mean that our model performs optimally, since it only adjusts to reward statistics in the environment without any optimization process. Not surprisingly, however, the change-detection Bayesian model, which was designed and tailored for superior performance in the PRL task, outperformed all other models and its performance reached to that of an omniscient observer (Figure 6A).

To test our model's robustness more rigorously, we also measured the performance in a "universe" where the level of uncertainty and volatility changed every few blocks of trials (see STAR Methods). The result of this simulation showed that there are certain ranges of learning rates that allow RL(1) and RL(2) to perform reasonably well in such a dynamic environment (Figures 6B and 6C). In contrast, our model was able to perform well over a wide range of parameter values (Figure 6E). This indicates that learning based on our proposed metaplasticity does not require fine-tuning to achieve a high level of performance, mainly because it can flexibly adjust to reward statistics in the environment.

We also examined how the model's behavior depends on the number of weak and strong meta-states, m (Figures 6D–6F). The model's performance was high for a wide range of parameter values (p_1 and q_1) and for different number of meta-states. This shows that even with a small number of meta-states the model can robustly perform the PRL task. Moreover, the maximal performance of the RMDP model matches that of optimal RL(2) and exceeds that of optimal RL(1), even though the RDMP model does not have separate transition probabilities for potentiation and depression events (Figure 6B, inset). These results indicate that RDMP can improve performance in dynamic/mixed environments. Moreover, in such environments, having more meta-states can slightly worsen the performance and restrict the range of parameters for which the model's performance is high (Figures 6D–6F). These effects occur because with a larger number of meta-states, synapses can more easily get "stuck" in deep meta-states for certain sets of parameters, which can reduce adaptability. Overall, our simulations illustrate that our model can perform reward integration in dynamic environments over a wide range of parameters using a small number of meta-states.

A recent study has reported that metaplasticity alone does not provide enough flexibility to capture learning under reward uncertainty (Iigaya, 2016). To achieve optimal behavior, Iigaya (2016) incorporated an additional network that computes expected and unexpected uncertainty over several different timescales to detect "surprise" on a specific timescale in order to update corresponding transition rates in the metaplastic network in

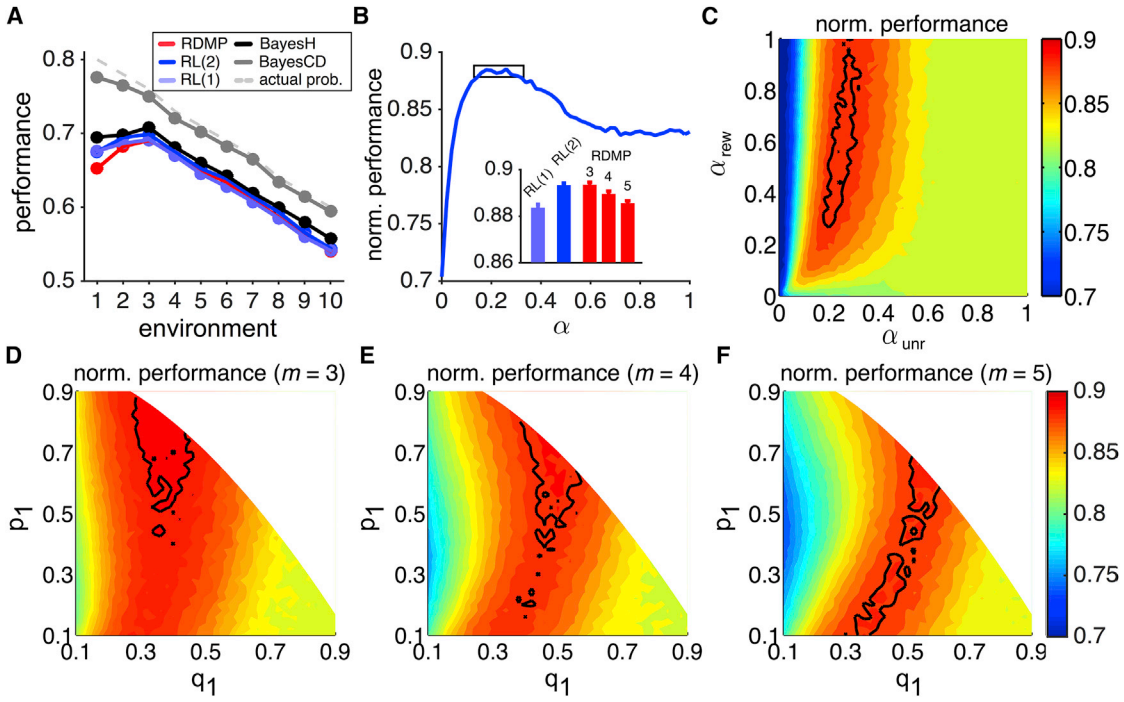

Figure 6. The RDMP Model Robustly Performs the PRL Task

(A) Performance of five different models in ten selected environments that require different optimal learning rates (BayesH: hierarchical Bayesian; BayesCD: change-detection Bayesian). The performance of the RDMP model is computed using one set of parameters in all environments, whereas the performance for the RL(2) and RL(1) models are based on the optimal learning rates chosen separately for each environment. The performance of the omniscient observer that knows the better option and chooses that option all the time is equal to the actual probability of reward assignment on the better option.

(B) Performance (normalized by the performance of the omniscient observer) of RL(1) in a universe with many different levels of uncertainty/volatility, as a function of the learning rate. The normalized performance of 0.7 corresponds to chance performance. The inset shows the optimal performance (±SD) of the RL(1), RL(2), and RDMP models with different number of meta-states (3, 4, and 5), computed by averaging top 2% performance in order to reduce noise. The rectangle indicates the top 2% performance.

(C) The performance of RL(2) in a universe with many different levels of uncertainty/volatility, as a function of the learning rates for rewarded and unrewarded trials. The black curves enclose the top 2% performance.

(D–F) The performance of RDMP in a universe with many different levels of uncertainty/volatility, as a function of the maximum transition probabilities, and for different numbers of meta-states. The white region indicates parameter values that could result in implausible transitions in the model (see STAR Methods).

an ad hoc fashion. However, Iigaya utilized the same metaplasticity architecture as the cascade model of Fusi et al. (2005), which was designed to preserve memory over long timescales. By contrast, our model is more general and includes transitions between meta-states not present in the cascade model (upward vertical arrows in Figure 2A). These additional metaplastic transitions can de-stabilize synapses without changing their efficacy.

To show that having these metaplastic transitions is critical for flexibility, we simulated the behavior of a single-parameter version of our model and the cascade model used by Iigaya (Figures S5A and S5B) in a universe with many different levels of uncertainty and volatility. Our model significantly outperformed the cascade model for most values of x, the single transition probability that determines the largest transition probability in both models (Figure S5C). Moreover, the difference in performance between the two models increased with a larger number of meta-states. Note that the chance level for the normalized performance is 0.7 due to the probabilistic nature of the PRL task. These results suggest that metaplastic transitions that can de-stabilize synapses without changing their efficacy are critical for achieving adaptability in reward integration under uncertainty. Because of these transitions, our model can more quickly switch its behavior after reversals and move toward a steady state.

Note that our main claim about the usefulness of metaplasticity is not contingent on exactly how transition probabilities depend on the level of meta-states (e.g., power law in Equation 2 of STAR Methods). Instead, we propose that any flexible RDMP model is suitable for adaptive learning if it exhibits a larger effective learning rate for reward assignment on the better than the one on the worse option at the steady state. Such a model of metaplasticity can perform reasonably well without being optimal, but more importantly, can capture the experimental data.

Neural Correlates of Reward Uncertainty

So far we showed that our model can robustly perform the PRL task while metaplastic synapses are adjusting to the uncertainty and volatility in the environment and that the model captures

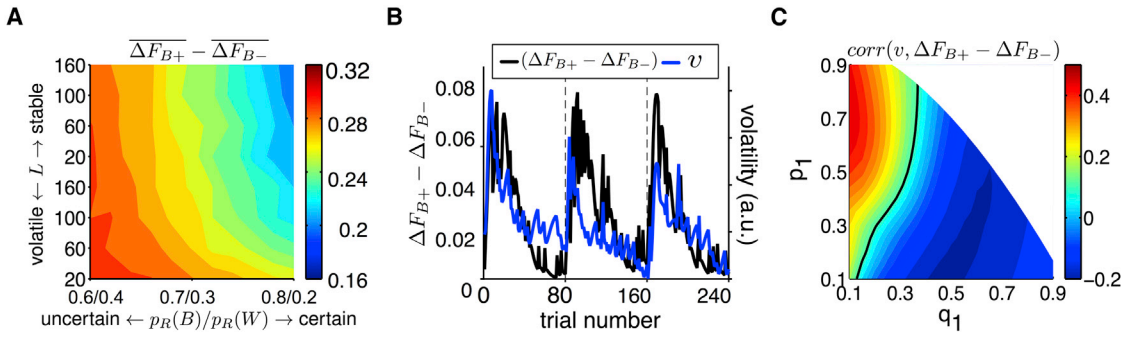

Figure 7. Neural Correlates of Estimated Volatility in the RDMP Model

(A) Plotted is the average value of the difference in the changes in synaptic strengths in the RDMP model, for different environments.

(B) The time course of the difference in the change in synaptic strengths in the RDMP model (blue), and of estimated volatility from the hierarchical Bayesian model (black) during three blocks of trials in the stable environment. For these simulations we used $q_1 = 0.2$ and $p_1 = 0.6$.

(C) The correlation coefficient between trial-by-trial estimate of $(\Delta F_{B+}(t) - \Delta F_{B-}(t))$ and estimated volatility by the hierarchical Bayesian model over a wide range of model's parameters (the maximum transition probabilities), during ten environments with different levels of volatility (block length). The black curve indicates parameter values for which the correlation is equal to 0.1.

important features of the experimental data better than all other competing models. Given that a signal related to volatility estimates in the hierarchical Bayesian model was identified in the anterior cingulate cortex (ACC) (Behrens et al., 2007), we next investigated whether there is any signal in our model that can be used as a proxy for volatility estimated by the hierarchical Bayesian model (since our model does not directly estimate volatility). The presence of such signals would explain how neural correlates of volatility could be detected even without any Bayesian computations of volatility.

In the hierarchical Bayesian model, volatility "v" determines the width of transition probability between two consecutive estimates of reward probability (see STAR Methods). Similarly, in our model, abs($\Delta F_{B+}(t)$) + abs($\Delta F_{B-}(t)$) determines the scale of changes between two consecutive estimates of reward probability. Because $\Delta F_{B-}(t)$ is always negative, the above quantity can be computed by $(\Delta F_{B+}(t) - \Delta F_{B-}(t))$. Indeed, we found that the average value of the difference in synaptic strength changes due to two possible reward assignments, $(\overline{\Delta F_{B+}} - \overline{\Delta F_{B-}})$, strongly depends on the uncertainty and volatility in a given environment (Figure 7A). With a candidate signal for the volatility in our model, we next asked whether this signal might be correlated with volatility estimated by the hierarchical Bayesian model.

We found that for certain model's parameters, the time course of $(\Delta F_{B+}(t) - \Delta F_{B-}(t))$ closely resembled that of estimated volatility from the hierarchical Bayesian model (Figure 7B). This difference can be estimated by the response of neurons that represent the latest change in the activity of value encoding neurons between consecutive trials. We found a significant trial-by-trial correlation between this approximation of $(\Delta F_{B+}(t) - \Delta F_{B-}(t))$ and estimated volatility over a wide range of model's parameters, and across many environments with different levels of uncertainty and volatility (Figure 7C). As a comparison, we also computed the correlation between estimated volatility and change in the value function in the RL(2) model $(\Delta V_{B+}(t) - \Delta V_{B-}(t))$. This correlation, however, was weak and only observed for very limited values of RL(2)'s parameters (Figure S6). These results show that without computing volatility explicitly, our model based on metaplasticity can generate a signal that correlates with estimated volatility, and therefore, might account for the signal observed in the ACC (Behrens et al., 2007).

Our model predicts differential responses when reward is assigned to the better and worse options as it progresses into a block of trials (compare $\Delta F_{B+}(t)$ and $\Delta F_{B-}(t)$ in Figure 3). Therefore, correlation between estimated volatility and changes in synaptic strength might differ depending on whether the better or worse option was rewarded. However, the trial-by-trial approximation for changes in the synaptic strength based on their latest value could result in a stronger correlation for trials on which reward was assigned to the better option merely because of the larger number of these trials. To avoid this confound, we also computed the correlation between the average time course of $\Delta F_{B+}(t)$ (or $\Delta F_{B-}(t)$) and estimated volatility based on the hierarchical Bayesian model (Figure S7). Indeed, we found a stronger correlation between these two measures for trials when reward was assigned to the better option. Therefore, in addition to providing a mechanistic account of the volatility signal observed in the ACC, our model also predicts that this signal should depend on which option reward is assigned to. This prediction can be tested in future experiments.

DISCUSSION

Adjustment of Learning to Reward Uncertainty

Adaptive decision making relies on estimating the reward values of objects or actions that have to be constantly updated, since those values can unpredictably change over time in an uncertain world (Bland and Schaefer, 2012; Courville et al., 2006; Mathys et al., 2011; O'Reilly, 2013). There are two problems at the heart of this estimation, depending on the model used to tackle reward under uncertainty. First, there is a tradeoff between having an accurate estimate of reward values and being able to quickly update those values due to changes in the environment (adaptability-precision tradeoff). Second, estimating uncertainty is very

challenging without a proper model of the environment, but such estimation is the foundation upon which alternative models of the environment could be built (Bland and Schaefer, 2012; Courville et al., 2006; O'Reilly, 2013).

Our model based on RDMP partially circumvents the first problem by adjusting learning based on reward statistics, and moreover, can generate a signal that can be used to build a model of the environment. More specifically, the model increases the effective learning rate on trials when reward is assigned to the better option and decreases the learning rate on trials when reward is assigned to the worse option, as the model experiences particular reward statistics. The difference between these learning rates increases as volatility or uncertainty decreases. These adjustments allow better integration of signal while ignoring noise and, thus, improve precision in detecting the more rewarding option. The same mechanism, however, causes the model to be initially slow in responding to real changes in the environment. Nevertheless, after receiving a few consecutive outcomes in the opposite direction of what the model has previously learned about the environment, synapses transition to more unstable meta-states allowing the model to become adaptable again. Interestingly, our model can significantly predict choice behavior better than optimal models during such sequences of trials.

A few studies have shown that the learning rates sharply increase and then decay when a change point in reward statistics occurs (Nassar et al., 2010; Diederen and Schultz, 2015). By adjusting to reward feedback, our model adopts two separate learning rates for reward assignments on the better and worse options. The learning rate increases over trials when the reward outcome supports the currently better option, whereas it decreases when the outcome supports the currently worse option. Although one of these changes is consistent with the results of the aforementioned studies, there are critical differences between the tasks used in those and our studies. In those studies, the subjects had to predict the value of a continuous variable and were provided with the error in their prediction on every trial. This task is very different from estimating reward probability based on binary feedback without a possibility to detect abrupt changes in the estimated quantity using a single reward outcome. Future studies are required to test whether separate learning rates also exists for estimation of continuous reward outcomes.

Neural Substrates of Adaptive Choice Under Uncertainty

A previous study on the neural substrates of uncertainty has shown that the BOLD signal in the ACC reflects volatility (unexpected uncertainty) and that variations in ACC signals are predictive of subject learning rates (Behrens et al., 2007). Here, we show that changes in the activity of model neurons endowed with RMDP can be used to estimate volatility. Interestingly, it has been suggested that the ACC projections to the locus coeruleus (LC) enable target neurons to signal unexpected uncertainty (Aston-Jones and Cohen, 2005), which is assumed to rely on the norepinephrine (NE) system (Preuschoff et al., 2011; Yu and Dayan, 2005). Therefore, our results show the feasibility of this mechanism assuming the presence of metaplastic synapses in the ACC, but more importantly, also suggest a plausible neural substrate for generating the observed uncertainty signal in the ACC. An analogous signal can be generated in our model simply by adding the absolute changes in value estimates when reward is assigned to the better option and when reward is assigned to the worse option, without having an explicit model of the environment required for a Bayesian estimation of volatility. Furthermore, our model predicts a stronger correlate of estimated volatility when reward is assigned to the better option.

A lesion study on the role of ACC in reward learning in dynamic environments has shown that ACC-lesioned animals were unable to sustain a response that yielded reward and displayed a reduction in the time constant of reward integration leading to impaired learning, especially when reward probabilities were low (Kennerley et al., 2006; Rushworth and Behrens, 2008). Our results suggest that metaplasticity allows more precise estimation of reward probability without a significant loss in adaptability. If we consider the ACC as a nexus for metaplasticity, we could assume that after losing a "metaplastic" evaluation system, ACC-lesioned monkeys would rely more on "non-metaplastic" evaluation systems (e.g., in basal ganglia), which are less capable of mitigating the adaptability-precision tradeoff. To handle volatility in the environment, the lesioned animals could increase their learning rates resulting in more noise in their estimation of reward value and, therefore, poorer performance in dynamic environments (Rushworth and Behrens, 2008). This impairment in performance would be more pronounced when the reward probability is low, resulting in less frequent reward feedback, which is consistent with the data (Kennerley et al., 2006). Therefore, our model predicts that lack of a metaplastic evaluation system (perhaps in ACC) should generally result in noisier behavior under reward uncertainty.

Although reward signal utilized in our model is generally believed to be transmitted by the neurotransmitter dopamine (DA) (Schultz, 2002), others have suggested dedicated neuromodulator systems for signaling different types of uncertainty (Bland and Schaefer, 2012; Yu and Dayan, 2005). Nevertheless, several pieces of evidence suggest that DA is a plausible neuromodulator for guiding reward integration under uncertainty. First, DA is the main neurotransmitter for signaling reward (Schultz, 2002), so any computations underlying reward uncertainty is likely to rely on DA-dependent plasticity. Second, DA affects neural processes at multiple timescales (Schultz, 2007). Interestingly, a recent study using a PRL task found that the activity of neurons in the ACC (which receives dopaminergic inputs) and not in the lateral habenula (which inhibits midbrain DA neurons) is strongly modulated by consecutive negative outcomes (Kawai et al., 2015). Finally, there is indirect experimental evidence for dopamine-dependent metaplasticity (Moussawi et al., 2009). Altogether, these pieces of evidence and our results suggest that computations underlying reward uncertainty could rely on DA-dependent metaplasticity in the ACC, though other neuromodulator systems might be involved for improving these computations further.

Relationship to Existing Models

Previous models for choice under reward uncertainty have proposed roughly three different mechanisms for learning from

reward feedback: (1) to determine the optimal learning rates based on optimization or on the level of uncertainty in a given environment (RL models) (Doya, 2002; Schweighofer and Doya, 2003; Preuschoff and Bossaerts, 2007); (2) to identify the correct model of the world, which includes the amount of uncertainty to properly incorporate reward feedback (Bayesian models) (Behrens et al., 2007; Courville et al., 2006; Daw et al., 2005, 2006; Hampton et al., 2006; Payzan-LeNestour and Bossaerts, 2011; Yu and Dayan, 2005); and (3) to detect changes in the environment to adapt learning accordingly (approximate Bayesian models) (Gallistel et al., 2014; Jang et al., 2015; Mathys et al., 2011; McGuire et al., 2014; Nassar et al., 2010, 2012; Wilson et al., 2013, 2014).

Having the correct model of the environment, the Bayesian models are able to outperform simple RL models, which have only limited access to the state of the environment (Behrens et al., 2007; Hampton et al., 2006; Jang et al., 2015; Payzan-LeNestour and Bossaerts, 2011). Alternatively, one could assign values to different states of the world and instead of re-learning the value of each action, only estimates the state of the world, which could greatly enhance behavioral adaptation (Wilson et al., 2014). Nevertheless, when certain assumptions are relaxed, the exact Bayesian inference becomes intractable and it is unclear what approximations subjects should make to perform the necessary computations (Courville et al., 2006, but see Nassar et al., 2010, 2012, and Wilson et al., 2013). Interestingly, most Bayesian models of choice under uncertainty assume a hierarchical structure for uncertainty (e.g., expected and unexpected uncertainty) on different timescales (Behrens et al., 2007; Mathys et al., 2011; McGuire et al., 2014; Payzan-LeNestour and Bossaerts, 2011; Wilson et al., 2013; Yu and Dayan, 2005) or assume the correct structure of the task at hand (Costa et al., 2015; Daw et al., 2002; Hampton et al., 2006; Jang et al., 2015). It is, however, unclear how the brain can separate different types of uncertainty or construct the proper model of the environment. Indeed, in the absence of any task instructions reflecting the structure of uncertainty, experimental results do not support the use of the Bayesian approach for dealing with uncertainty (Payzan-LeNestour and Bossaerts, 2011).

In a recent work, Gallistel and colleagues showed that human behavior during estimation of the probability of a binary outcome, which unpredictably changes over time, cannot be explained by delta-rule models (Gallistel et al., 2014). They instead suggested a new Bayesian model based on change-point estimation and with evidence-triggered updating. In their model, the estimate of reward probability is updated only if a change is detected and, if so, a new estimate of reward probability can be made depending on the location of the detected change. Interestingly, a recent modeling study has shown that increased responsiveness to change-points can be instantiated by pauses in tonically active interneurons in the striatum enabling the modulation of learning rate by reward uncertainty (Franklin and Frank, 2015). Although we did not incorporate a change-detection mechanism, such a mechanism would only improve the performance of our model (Gallistel et al., 2001; McGuire et al., 2014). Nevertheless, we suggest that via RDMP, reward statistics itself can directly affect the level of plasticity and thus determine both learning and its adjustment according to reward history and statistics.

A series of studies have provided approximations for full Bayesian models of learning under uncertainty, and demonstrated that these approximate models provide a better fit to experimental data than full Bayesian models (Mathys et al., 2011; Nassar et al., 2010, 2012; Wilson et al., 2013). Interestingly, some of these models can be mapped onto learning based on delta-rules and have a hierarchical structure where updates in one level depends on estimate in another level (Mathys et al., 2011; Wilson et al., 2013). Our model also has a hierarchical structure, but the rate at which synapses transition between a given set of meta-states is fixed and not controlled by information in another level. Moreover, unlike the aforementioned approximate Bayesian models, learning in our model depends on a binary reward signal and not on reward prediction error. The multiple time constants for updates and hierarchical structures in aforementioned models might rely on the observed reservoir of reward memory time constants (Bernacchia et al., 2011) or could be a reflection of metaplasticity proposed here, since metaplastic synapses contain different internal timescales.

Finally, a recent study claimed that metaplastic synapses are limited in capturing abrupt changes in the environment and suggested that a surprise detection system is necessary for adaptive integration of reward feedback (Iigaya, 2016). Here, we show that the metaplastic architecture utilized in that study is quite inflexible, since it was originally designed for keeping memory over long timescales (Fusi et al., 2005), leading to an incorrect conclusion that metaplasticity cannot provide the needed solution. Although our model does not have a mechanism dedicated for detecting abrupt changes in the environment and thus is suboptimal in shifting behavior after reversals, it captures subjects' behavior better than competing optimal models during the same exact shifts.

Computational Power of Metaplasticity

Although there is a large body of experimental evidence for metaplasticity (Abraham, 2008), metaplasticity has been mainly used to explain low-level phenomena such as changes in the threshold for induction of plasticity due to prior synaptic activity (Yger and Gilson, 2015). The contribution of metaplasticity to behavior is mostly unknown and our study is the first to provide direct behavioral evidence for it. Generally speaking, any model that exhibits synaptic changes without changes in synaptic efficacy can capture a form of metaplasticity (Fusi et al., 2005). Our model incorporates modifications to dopamine-dependent Hebbian learning rules (Reynolds and Wickens, 2002; Soltani and Wang, 2006) that depend on activity in the preceding trials, and thus extend the experimentally observed effects of metaplasticity to the realm of reward-dependent learning for which there is some indirect experimental evidence (Moussawi et al., 2009). Our goal here was to show that, in principle, metaplasticity could provide a plausible solution to an important problem in value-based learning. The novelty of our proposal is that it provides a plausible and robust low-level (i.e., synaptic level) mechanism for a seemingly high-level cognitive function, entirely based on metaplasticity.

Metaplastic synapses are strong computational tools because of the many possible transitions they contain, many of which do not change synaptic efficacy, making meta-states similar to hidden layers in Markov chains (Rabiner, 1989). The space of possible metaplastic models is immense and our exploration of metaplastic models suitable for learning in dynamic environments has revealed an important component of these models: the ability to destabilize weak (strong) synapses while stabilizing strong (weak) synapses on potentiation (depression) events. Although our proposed metaplasticity architecture still needs to be tested in future experiments, this architecture allows metaplastic models to be adaptable without significantly increasing noise, and thus mitigating the adaptability-precision tradeoff. Together, our work highlights the overlooked power of metaplastic synaptic mechanisms for solving complex cognitive problems (Mongillo et al., 2008).

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - Behavioral task
  - Computational model
  - Learning rule
  - Model simulations
  - Computation of the effective learning rates
  - Alternative models
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Fitting of experimental data and data analysis

SUPPLEMENTAL INFORMATION

Supplemental Information includes seven figures and supplemental data and can be found with this article online at http://dx.doi.org/10.1016/j.neuron.2017.03.044.

AUTHOR CONTRIBUTIONS

S.F., P.K., and A.S. designed the model. C.H.D. and D.L. designed the experiment. S.F., P.K., and A.S. performed model simulations and analyzed the data. C.H.D. conducted the experiment. C.H.D., S.F., D.L., H.S., and A.S. analyzed and interpreted the experimental data. D.L. and A.S wrote the manuscript and all other authors contributed to revising the manuscript.

ACKNOWLEDGMENTS

We would like to Zohra Aslami, Brad Duchaine, Clara Guo, and Katherine Rowe for helpful comments on the manuscript. This work was supported by National Science Foundation (EPSCoR Award #1632738) and Neukom Institute CompX Grant to A.S., and National Institutes of Health (R01 DA029330 and R01 MH108629 to D.L.).

Received: April 30, 2016
Revised: September 2, 2016
Accepted: March 29, 2017
Published: April 19, 2017

REFERENCES

Abraham, W.C. (2008). Metaplasticity: tuning synapses and networks for plasticity. Nat. Rev. Neurosci. *9*, 387–399.

Abraham, W.C., and Bear, M.F. (1996). Metaplasticity: the plasticity of synaptic plasticity. Trends Neurosci. *19*, 126–130.

Aston-Jones, G., and Cohen, J.D. (2005). An integrative theory of locus coeruleus-norepinephrine function: adaptive gain and optimal performance. Annu. Rev. Neurosci. *28*, 403–450.

Behrens, T.E.J., Woolrich, M.W., Walton, M.E., and Rushworth, M.F.S. (2007). Learning the value of information in an uncertain world. Nat. Neurosci. *10*, 1214–1221.

Bernacchia, A., Seo, H., Lee, D., and Wang, X.-J. (2011). A reservoir of time constants for memory traces in cortical neurons. Nat. Neurosci. *14*, 366–372.

Bland, A.R., and Schaefer, A. (2012). Different varieties of uncertainty in human decision-making. Front. Neurosci. *6*, 85.

Cools, R., Barker, R.A., Sahakian, B.J., and Robbins, T.W. (2001). Enhanced or impaired cognitive function in Parkinson's disease as a function of dopaminergic medication and task demands. Cereb. Cortex *11*, 1136–1143.

Costa, V.D., Tran, V.L., Turchi, J., and Averbeck, B.B. (2015). Reversal learning and dopamine: a bayesian perspective. J. Neurosci. *35*, 2407–2416.

Courville, A.C., Daw, N.D., and Touretzky, D.S. (2006). Bayesian theories of conditioning in a changing world. Trends Cogn. Sci. *10*, 294–300.

Daw, N.D., Kakade, S., and Dayan, P. (2002). Opponent interactions between serotonin and dopamine. Neural Netw. *15*, 603–616.

Daw, N.D., Niv, Y., and Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nat. Neurosci. *8*, 1704–1711.

Daw, N.D., O'Doherty, J.P., Dayan, P., Seymour, B., and Dolan, R.J. (2006). Cortical substrates for exploratory decisions in humans. Nature *441*, 876–879.

Diederen, K.M., and Schultz, W. (2015). Scaling prediction errors to reward variability benefits error-driven learning in humans. J. Neurophysiol. *114*, 1628–1640.

Donahue, C.H., and Lee, D. (2015). Dynamic routing of task-relevant signals for decision making in dorsolateral prefrontal cortex. Nat. Neurosci. *18*, 295–301.

Doya, K. (2002). Metalearning and neuromodulation. Neural Netw. *15*, 495–506.

Frank, M.J., Moustafa, A.A., Haughey, H.M., Curran, T., and Hutchison, K.E. (2007). Genetic triple dissociation reveals multiple roles for dopamine in reinforcement learning. Proc. Natl. Acad. Sci. USA *104*, 16311–16316.

Frank, M.J., Doll, B.B., Oas-Terpstra, J., and Moreno, F. (2009). Prefrontal and striatal dopaminergic genes predict individual differences in exploration and exploitation. Nat. Neurosci. *12*, 1062–1068.

Franklin, N.T., and Frank, M.J. (2015). A cholinergic feedback circuit to regulate striatal population uncertainty and optimize reinforcement learning. eLife *4*, e12029.

Fusi, S., Drew, P.J., and Abbott, L.F. (2005). Cascade models of synaptically stored memories. Neuron *45*, 599–611.

Gallistel, C.R., Mark, T.A., King, A.P., and Latham, P.E. (2001). The rat approximates an ideal detector of changes in rates of reward: implications for the law of effect. J. Exp. Psychol. Anim. Behav. Process. *27*, 354–372.

Gallistel, C.R., Krishan, M., Liu, Y., Miller, R., and Latham, P.E. (2014). The perception of probability. Psychol. Rev. *121*, 96–123.

Hampton, A.N., Bossaerts, P., and O'Doherty, J.P. (2006). The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans. J. Neurosci. *26*, 8360–8367.

Hulme, S.R., Jones, O.D., and Abraham, W.C. (2013). Emerging roles of metaplasticity in behaviour and disease. Trends Neurosci. *36*, 353–362.

Iigaya, K. (2016). Adaptive learning and decision-making under uncertainty by metaplastic synapses guided by a surprise detection system. eLife *5*, e18073.

Jang, A.I., Costa, V.D., Rudebeck, P.H., Chudasama, Y., Murray, E.A., and Averbeck, B.B. (2015). The role of frontal cortical and medial-temporal lobe

brain areas in learning a bayesian prior belief on reversals. J. Neurosci. *35*, 11751–11760.

Kawai, T., Yamada, H., Sato, N., Takada, M., and Matsumoto, M. (2015). Roles of the Lateral Habenula and Anterior Cingulate Cortex in Negative Outcome Monitoring and Behavioral Adjustment in Nonhuman Primates. Neuron *88*, 792–804.

Kennerley, S.W., Walton, M.E., Behrens, T.E.J., Buckley, M.J., and Rushworth, M.F.S. (2006). Optimal decision making and the anterior cingulate cortex. Nat. Neurosci. *9*, 940–947.

Krugel, L.K., Biele, G., Mohr, P.N.C., Li, S.-C., and Heekeren, H.R. (2009). Genetic variation in dopaminergic neuromodulation influences the ability to rapidly and flexibly adapt decisions. Proc. Natl. Acad. Sci. USA *106*, 17951–17956.

Mathys, C., Daunizeau, J., Friston, K.J., and Stephan, K.E. (2011). A bayesian foundation for individual learning under uncertainty. Front. Hum. Neurosci. *5*, 39.

McGuire, J.T., Nassar, M.R., Gold, J.I., and Kable, J.W. (2014). Functionally dissociable influences on learning rate in a dynamic environment. Neuron *84*, 870–881.

Mongillo, G., Barak, O., and Tsodyks, M. (2008). Synaptic theory of working memory. Science *319*, 1543–1546.

Moussawi, K., Pacchioni, A., Moran, M., Olive, M.F., Gass, J.T., Lavin, A., and Kalivas, P.W. (2009). N-Acetylcysteine reverses cocaine-induced metaplasticity. Nat. Neurosci. *12*, 182–189.

Müller-Dahlhaus, F., and Ziemann, U. (2015). Metaplasticity in human cortex. Neuroscientist *21*, 185–202.

Nassar, M.R., Wilson, R.C., Heasly, B., and Gold, J.I. (2010). An approximately Bayesian delta-rule model explains the dynamics of belief updating in a changing environment. J. Neurosci. *30*, 12366–12378.

Nassar, M.R., Rumsey, K.M., Wilson, R.C., Parikh, K., Heasly, B., and Gold, J.I. (2012). Rational regulation of learning dynamics by pupil-linked arousal systems. Nat. Neurosci. *15*, 1040–1046.

Niv, Y., Edlund, J.A., Dayan, P., and O'Doherty, J.P. (2012). Neural prediction errors reveal a risk-sensitive reinforcement-learning process in the human brain. J. Neurosci. *32*, 551–562.

O'Reilly, J.X. (2013). Making predictions in a changing world-inference, uncertainty, and learning. Front. Neurosci. *7*, 105.

Payzan-LeNestour, E., and Bossaerts, P. (2011). Risk, unexpected uncertainty, and estimation uncertainty: Bayesian learning in unstable settings. PLoS Comput. Biol. *7*, e1001048.

Preuschoff, K., and Bossaerts, P. (2007). Adding prediction risk to the theory of reward learning. Ann. N Y Acad. Sci. *1104*, 135–146.

Preuschoff, K., 't Hart, B.M., and Einhäuser, W. (2011). Pupil dilation signals surprise: evidence for noradrenaline's role in decision making. Front. Neurosci. *5*, 115.

Rabiner, L.R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE *77*, 257–286.

Reynolds, J.N., and Wickens, J.R. (2002). Dopamine-dependent plasticity of corticostriatal synapses. Neural Netw. *15*, 507–521.

Rudebeck, P.H., Saunders, R.C., Prescott, A.T., Chau, L.S., and Murray, E.A. (2013). Prefrontal mechanisms of behavioral flexibility, emotion regulation and value updating. Nat. Neurosci. *16*, 1140–1145.

Rushworth, M.F.S., and Behrens, T.E.J. (2008). Choice, uncertainty and value in prefrontal and cingulate cortex. Nat. Neurosci. *11*, 389–397.

Rygula, R., Walker, S.C., Clarke, H.F., Robbins, T.W., and Roberts, A.C. (2010). Differential contributions of the primate ventrolateral prefrontal and orbitofrontal cortex to serial reversal learning. J. Neurosci. *30*, 14552–14559.

Schultz, W. (2002). Getting formal with dopamine and reward. Neuron *36*, 241–263.

Schultz, W. (2007). Multiple dopamine functions at different time courses. Annu. Rev. Neurosci. *30*, 259–288.

Schweighofer, N., and Doya, K. (2003). Meta-learning in reinforcement learning. Neural Netw. *16*, 5–9.

Soltani, A., and Wang, X.-J. (2006). A biophysically based neural model of matching law behavior: melioration by stochastic synapses. J. Neurosci. *26*, 3731–3744.

Soltani, A., and Wang, X.-J. (2010). Synaptic computation underlying probabilistic inference. Nat. Neurosci. *13*, 112–119.

Soltani, A., Lee, D., and Wang, X.-J. (2006). Neural mechanism for stochastic behaviour during a competitive game. Neural Netw. *19*, 1075–1090.

Swainson, R., Rogers, R.D., Sahakian, B.J., Summers, B.A., Polkey, C.E., and 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*, 596–612.

Wilson, R.C., Nassar, M.R., and Gold, J.I. (2013). A mixture of delta-rules approximation to bayesian inference in change-point problems. PLoS Comput. Biol. *9*, e1003150.

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

Yger, P., and Gilson, M. (2015). Models of metaplasticity: a review of concepts. Front. Comput. Neurosci. *9*, 138.

Yu, A.J., and Dayan, P. (2005). Uncertainty, neuromodulation, and attention. Neuron *46*, 681–692.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Experimental Models: Organisms/Strains		
Rhesus macaque (Mucacca mulatta)	Yale University	N/A
Software and Algorithms		
MATLAB	MathWorks	https://www.mathworks.com/products/matlab.html

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Dr. Alireza Soltani (soltani@dartmouth.edu).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Two male rhesus monkeys (age 5.8 and 5.1 years) were used. One monkey had been previously trained on a manual joystick task before this experiment and the other monkey had not been used for any prior experiments. Eye movements were monitored using an infrared eye tracker (ET49, Thomas Recording, Germany). All experimental procedures were approved by the Institutional Animal Care and Use Committee (IACUC) at Yale University. More details have been reported previously (Donahue and Lee, 2015).

METHOD DETAILS

Behavioral task

The animals were trained on a modified probabilistic reversal learning (PRL) task. In this task, two color targets (red and green) appeared on the screen after the animals fixated on a white square (c.f. Figure 1A in Donahue and Lee, 2015). After a 500-ms interval, a set of yellow tokens was presented around each target, indicating the magnitude of potential reward on a given target. Importantly, on each trial, one of the target colors was associated with a high reward probability (80%), and the other was associated with a low reward probability (20%). These reward probabilities were fixed within a block of trials and alternated across blocks of 20 or 80 trials ($L = 20$, or 80) so that the animals had to learn them through experience. The central fixation cue was extinguished following a random interval (ranging from 500 ms to 1200 ms) after which the animals were free to shift their gaze toward one of the two color targets. The animals received visual feedback after fixating the chosen target for 500 ms. A red or green ring around the chosen target indicated that the animals would be rewarded (after another 500 ms), while a gray or blue ring indicated that they were not to be rewarded. On trials where the animals were rewarded, they received the amount of apple juice associated with the chosen target. Each token corresponded to one drop of juice (0.1 mL). The reward magnitudes associated with each target color were drawn from the following ten possible pairs: {1,1}, {1,2}, {1,4}, {1,8}, {2,1}, {2,4}, {4,1}, {4,2}, {4,4}, {8,1}. Each magnitude pair was counter-balanced across target locations so that reward magnitude did not provide any information about the location of reward. We did not find any systematic differences in either animal's behavior and therefore, we combined the data from both monkeys in the analyses (a total of 118 sessions and 66,148 trials). More details about the task and behaviors of the animals have been reported previously (Donahue and Lee, 2015).

Computational model

We constructed a model to simulate the choice behavior during a PRL task or its modified version (see above). In the regular PRL task, the subject selects between two alternative options (e.g., red and green targets) that deliver reward probabilistically (Figure 1A). The probability of reward on the green and red options, $p_R(g)$ and $p_R(r)$, are complementary, but these probabilities switch (i.e., reverse) after a certain number of trials referred to as block length, L. We refer to the option with a larger and smaller reward probability on a given block as the better and worse option, respectively. The model consists of the value-encoding and decision-making circuit. The value-encoding circuit contains two pools of value-encoding neurons representing the reward value of the two options. These neurons receive their inputs through a set of metaplastic synapses that estimate the probability of reward from the two options via a reward-dependent learning rule based on metaplasticity (see Learning rule). The decision-making circuit uses the output of value-encoding neurons in order to make a decision on each trial.

We assumed that metaplastic synapses have multiple meta-states associated with each of the two levels of synaptic efficacy: weak and strong (Figure 2A). Although for simplicity we assumed binary values for synaptic efficacy, our results also hold for the

case where there are multiple levels of efficacies. Depending on model's choice and reward outcome on each trial, metaplastic synapses could transition between meta-states with similar or different synaptic efficacy (see Learning rule). This property allows us to simulate metaplasticity, or changes in the synaptic state that influence future synaptic changes without any observable change in the efficacy of synaptic transmission (Abraham and Bear, 1996).

The output of value-encoding neurons associated with a given option reflects the overall synaptic efficacy of metaplastic synapses onto those neurons. We quantified the fractions of synapses in a given pool (e.g., associated with the green target) that are in different weak and strong meta-states as $f_{gi-}(t)$ and $f_{gi+}(t)$, respectively (g represents green, i represents the meta-state level, and t represents the trial number). Because there are two levels of synaptic efficacy, the overall synaptic efficacy for each set of metaplastic synapses can be quantified as the overall fraction of synapses in strong meta-states. This fraction, which we refer to as the 'synaptic strength', is equal to the sum of the fraction of synapses in each set that are in any strong meta-states, $F_{g+}(t)=\sum_{i=1}^{m} f_{gi+}(t)$ and $F_{r+}(t)=\sum_{i=1}^{m} f_{ri+}(t)$, where m is the number of meta-states (weak or strong).

As we have shown before, the decision on every trial only depends on the overall difference in the output of the two value-encoding pools (Soltani and Wang, 2006, 2010; Soltani et al., 2006). This difference is proportional to the difference in the synaptic strengths of the two pools. Therefore, the decision on each trial is determined stochastically with a probability:

$$P_g(t)=\frac{1}{1+\exp\left(-\frac{(F_{g+}(t)-F_{r+}(t))}{\sigma}\right)} \quad \text{(Eq. 1)}$$

where $P_g(t)$ is the probability of choosing the green target on trial t, and σ determines the stochasticity in choice. We set σ equal to 0.1 which resulted in variability in decision making comparable to the level observed experimentally (Donahue and Lee, 2015).

Learning rule

We assumed that the transition probability between different meta-states becomes smaller for more stable (or 'deeper') meta-states in an exponential fashion. We adopted this specific formulation of RDMP for simplicity to be able to explore its behavior over a range of parameters. However, we note that this model provides one of many of possible solutions for performing learning under reward uncertainty successfully. In this formulation, the transition probabilities associated with each meta-state is determined by a power law equation:

$$q_i^+ = q_1^{+\,((m-2)\times i+1)/m-1},\ q_i^- = q_1^{-\,((m-2)\times i+1)/m-1} \text{ for } 2\le \mathrm{i} \le \mathrm{m}$$

$$p_i^+ = \left(p_1^+\right)^i,\ p_i^- = \left(p_1^-\right)^i \text{ for } 2\le \mathrm{i} < m \quad \text{(Eq. 2)}$$

where m is the number of meta-states, q_1^+ is the probability for transitions between the most unstable weak to the most unstable strong meta-state, q_i^+ (for $i>1$) is the probability for the transitions from weak meta-state (i+1) to the most unstable strong meta-state (S_1), and p_i^+ is the probability for the transitions between the weak meta-states (i+1) and i, and between strong meta-states i and (i+1) (Figure 2A). Similarly, q_1^- is the probability for transitions between the most unstable strong to the most unstable weak meta-state, q_i^- (for $i>1$) is the probability for the transitions from strong meta-state (i+1) to the most unstable weak meta-state (W_1), and p_i^- is the probability for the transitions between the strong meta-states (i+1) and i, and between weak meta-states i and (i+1).

For simplicity, here we assumed equal transition probabilities for potentiation and depression events: $q_1^+ = q_1^- = q_1$ and $p_1^+ = p_1^- = p_1$, where q_1 and p_1 are the transition probability from and to the most unstable meta-states. Unless otherwise mentioned, the model simulations were done using $q_1 = 0.4$ and $p_1 = 0.3$ with four levels of meta-states ($m = 4$). Note that $p_1 = 0$ corresponds to a model without upward or downward transitions between meta-states and is equivalent to the RL(1) model with the learning rate equal to q_1 (i.e., a model without metaplasticity). Moreover, in our formulation, $m = 2$ model is also equivalent to the RL(1) model since $q_2 = q_1$ for $m = 2$. We assumed a large number of metaplastic synapses for each set of neurons representing or estimating the reward probability and thus, used a mean-field approach to simulate the change in synaptic efficacy for sets of synapses associated with alternative options (e.g., red and green targets).

The changes in the synaptic states on each trial depend on the model's choice and reward outcome. However, because during the PRL task the reward is assigned to one of the two options, the location of reward can be inferred on each trial and used to learn. Therefore, here we have assumed that the reward assignment on each trial can determine the direction of change in the synaptic efficacy and learning. More specifically, if reward is assigned to the green target on a given trial (independently of what is selected), synapses associated with the green target are potentiated whereas synapses associated with the red target are depressed. Similarly, if reward is assigned to the red target on a given trial (independently of what is selected), synapses associated with the red target are potentiated whereas synapses associated with the green target are depressed. This 'coupled' learning rule leads to the two sets of complementary synaptic strengths ($F_{g+}(t)=1-F_{r+}(t)$).

On potentiation events, synapses occupying weak meta-states stochastically (i.e., with certain probabilities) transition to less stable weak meta-states while synapses occupying strong meta-states stochastically transition to more stable strong meta-states (golden arrows in Figure 2A). Moreover, synaptic efficacy could increase by synapses occupying weak meta-states making transitions to the most unstable strong meta-state (S_1). Therefore, on trials when the set of synapses associated with the green target is

potentiated, the fractions of synapses in different meta-states are updated as the following (index g (r) in f_{gi} (f_{ri}) is dropped for better readability):

$$f_{1+} \rightarrow f_{1+} + \sum_{j=1}^{m} q_j f_{j-} - p_1 f_{1+} \quad \text{(Eq. 3)}$$

$$f_{1-} \rightarrow f_{1-} - q_1 f_{1-} + p_1 f_{2-}$$

$$f_{i+} \rightarrow f_{i+} + p_{i-1} f_{(i-1)+} - p_i f_{i+}, \text{ for } 1 < i < m$$

$$f_{i-} \rightarrow f_{i-} - q_i f_{i-} - p_{i-1} f_{i-} + p_i f_{(i+1)-}, \text{ for } 1 < i < m$$

$$f_{m+} \rightarrow f_{m+} + p_{m-1} f_{(m-1)+}$$

$$f_{m-} \rightarrow f_{m-} - q_m f_{m-} - p_{m-1} f_{m-}$$

Importantly, because of the stochastic nature of synaptic transitions, potentiation events may not change synaptic efficacy of some synapses.

On depression events, similar transitions happen but in the opposite direction causing weak (respectively, strong) synapses to become more (respectively, less) stable and making strong meta-states transition to the most unstable weak meta-state (W_1) and therefore, reducing synaptic efficacy (cyan arrows in Figure 2A). Therefore, on trials when this set of synapses is depressed, these fractions are updated as the following:

$$f_{1-} \rightarrow f_{1-} + \sum_{j=1}^{m} q_j f_{j+} - p_1 f_{1-} \quad \text{(Eq. 4)}$$

$$f_{1+} \rightarrow f_{1+} - q_1 f_{1+} + p_1 f_{2+}$$

$$f_{i-} \rightarrow f_{i-} + p_{i-1} f_{(i-1)-} - p_i f_{i-}, \text{ for } 1 < i < m$$

$$f_{i+} \rightarrow f_{i+} - q_i f_{i+} - p_{i-1} f_{i+} + p_i f_{(i+1)+}, \text{ for } 1 < i < m$$

$$f_{m-} \rightarrow f_{m-} + p_{m-1} f_{(m-1)-}$$

$$f_{m+} \rightarrow f_{m+} - q_m f_{m+} - p_{m-1} f_{m+}$$

Note that if both q_1 and p_1 are large, the Equations 2-4 may result in negative values for the fraction of synapses in certain meta-states. Therefore, we limited model's parameters to avoid such implausible transitions (white regions in Figures 6D–6F and 7C, and Figure S7).

Based on Equations 3-4, changes in the synaptic strength for synapses associated with the better option when reward was assigned to that option ($\Delta F_{B+}(t)$) or the alternative option ($\Delta F_{B-}(t)$) are equal to:

$$\Delta F_{B+}(t) = \sum_{j=1}^{m} q_j f_{Bj-}(t) \quad \text{(Eq. 5)}$$

$$\Delta F_{B-}(t) = -\sum_{j=1}^{m} q_j f_{Bj+}(t)$$

where $f_{Bj+}$and f_{Bj-} are the fraction of synapses associated with the better option which are in the strong and weak meta-state j, respectively. Because of the coupled nature of the learning rule, changes in the synaptic strength for synapses associated with the worse option, ΔF_{W+} and ΔF_{W-}, are the mirror image of those for the better option; $\Delta F_{W+} = -\Delta F_{B-}$ and $\Delta F_{W-} = -\Delta F_{B+}$. Finally, we defined the model's overall response to both types of reward feedback as the overall change in the synaptic strength for synapses associated with the better option as the weighted average of ΔF_{B+} and ΔF_{B-}:

$$\Delta F(t) = p_R(B)\ \Delta F_{B+}(t) + p_R(W)\Delta F_{B-}(t) \quad \text{(Eq. 6)}$$

where $p_R(B)$ and $p_R(W)$ are the probability of reward on the better and worse options in a given block, respectively.

Model simulations

To test the adjustments of our model to reward statistics in the environment, we simulated behavior in ten different environments requiring very different learning rates based on a simple reinforcement learning model (see Alternative models). The environment 1 to 10 (Figure 1C) are defined with the following parameters: $p_R(B)/p_R(W)$ = [0.6/0.4, 0.62/0.38, 0.65/0.35,0.67/0.33, 0.69/0.31, 0.71/0.29, 0.73/0.27, 0.76/0.24, 0.78/0.28, 0.8/0.2] and L = 200, 180, 160, 140, 120, 100, 80, 60, 40, and 20, where $p_R(B)$ and

$p_R(W)$ are the probability of reward on the better and worse options, respectively, and L is the block length. To further test robustness of metaplasticity, we also measured the performance in a 'universe' where the level of uncertainty changes more gradually across blocks. More specifically, the reward probability on the better option (i.e., the option with a higher reward probability) could change between 0.6 and 0.8 with a step size of 0.05 (the probability for the worse option was equal to 1 minus this number) while the block length could vary between 20, 50, 100, and 200 trials. A universe contains environments defined by all possible combinations of above probabilities and block lengths (each environment lasted for 2000 trials before changing to another environment). All other models were tested on a similar variable environment, and the results are based on average from ten randomly generated universes. For simulations presented in Figure 7C and Figures S6B and S7, we used a set of ten environments with reward probability equal to 0.8/0.2 and block length L = 20, 40, 60, 80, 100, 120, 140, 160, 180, 200.

Computation of the effective learning rates

In our model, learning is determined by reward history and therefore, changes over time. In order to capture the change in learning over time, we computed the 'effective' learning rates when reward was assigned to the better or worse option on a given block of trials. More specifically, the effective learning rate when reward was assigned to the better option on trial t after a reversal, $K_{B+}(t)$, was defined as the overall increase in the efficacy of metaplastic synapses associated with that option divided by the total fraction of those synapses in weak meta-states (using Equation 5),

$$K_{B+}(t) = \left(\sum_{j=1}^{m} q_j\, f_{Bj-}(t)\right) \Big/ \left(\sum_{j=1}^{m} f_{Bj-}(t)\right) = \Delta F_{B+}(t) \Big/ \left(\sum_{j=1}^{m} f_{Bj-}(t)\right) \quad \text{(Eq. 7a)}$$

Similarly, the effective learning rate when reward was assigned to the worse option, $K_{B-}(t)$, was defined as the overall decrease in the efficacy of respective metaplastic synapses divided by the total fraction of those synapses in strong meta-states,

$$K_{B-}(t) = \left(\sum_{j=1}^{m} q_j\, f_{Bj+}(t)\right) \Big/ \left(\sum_{j=1}^{m} f_{Bj+}(t)\right) = -\Delta F_{B-}(t) \Big/ \left(\sum_{j=1}^{m} f_{Bj+}(t)\right) \quad \text{(Eq. 7b)}$$

We refer to these ratios as the ''effective'' learning rates since they are analogous to the learning rates in a corresponding RL model at a given point in a block.

Note that the effective learning rates for synapses associated with the worse option were the mirror image of those for the better option. This is due to the coupled learning rule adopted for the PRL task, where reward is assigned to one of the two options, entailing that the two sets of synapses associated with the two options are updated in the opposite direction of each other on every trial. Relaxing the coupled learning rule results in two separate sets of learning rates for the two options, a possibility not considered further in the present study.

We also computed the effective learning rate on rewarded and unrewarded trials, $K_{rew}(t)$ and $K_{unr}(t)$, by simply averaging the effective learning rates based on choice and outcome on a given trial:

$$K_{rew}(t) = P_B(t) \times p_R(B) \times K_{B+}(t) + (1 - P_B(t)) \times p_R(W) \times K_{B-}(t) \quad \text{(Eq. 8)}$$

$$K_{unr}(t) = P_B(t) \times p_R(W) \times K_{B-}(t) + (1 - P_B(t)) \times p_R(B) \times K_{B+}(t)$$

where $P_B(t)$ is the probability of choosing the better option on trial t after a reversal, and $p_R(B)$ and $p_R(W)$ are the probability of reward on the better and worse options in a given block, respectively.

Alternative models

The model referred as RL(1) is a simple RL model based on the reward prediction error (RPE) and has two parameters: a single learning rate (α), and a temperature that determines the amount of stochasticity in decision-making process. In this model, the two options (red and green targets) are assigned with value functions V_g and V_r and the choice is determined based on the logistic function of the difference between these two values:

$$P_g(t) = \frac{1}{1 + \exp\left(-\left(V_g(t) - V_r(t)\right)/\sigma\right)} \quad \text{(Eq. 9)}$$

where σ determines the amount of stochasticity in choice. After every trial, the value functions are updated based on the reward assignment (assuming a 'coupled' learning rule):

$$V_g(t+1) = V_g(t) + \alpha\left(r(t) - V_g(t)\right),\ \mathrm{r(t)} = 1\,(0) \text{ for reward assigned to green (red)}$$

$$V_r(\mathrm{t}) = 1\text{-}V_g(\mathrm{t}) \quad \text{(Eq. 10)}$$

The model referred as RL(2) is a simple RL model based on the RPE and two separate learning rates for rewarded and unrewarded trials (α_{rew} and α_{unr}) instead of one as in RL(1). Therefore, the value functions are updated as:

$$V_g(t+1) = V_g(t) + \alpha_{rew}\left(1 - V_g(t)\right), \text{ reward assigned to green and green selected}$$

$$V_g(t+1) = V_g(t) + \alpha_{unr}\left(1 - V_g(t)\right), \text{ reward assigned to green but red selected}$$

$$V_g(t+1) = V_g(t) - \alpha_{rew}V_g(t), \text{ reward assigned to red and red selected}$$

$$V_g(t+1) = V_g(t) - \alpha_{unr}V_g(t), \text{ reward assigned to red but green selected} \quad \text{(Eq. 11)}$$

The decision rule was similar to RL(1). Unless otherwise mentioned, we used $\alpha_{rew} = 0.4$, $\alpha_{unr} = 0.2$, and $\sigma = 0.1$ for all RL(1) and RL(2) simulations.

The hierarchical Bayesian model is similar to the model presented in Behrens et al. (Behrens et al., 2007). Briefly, this model assumes a three-layer hierarchical structure for changes in reward probability over time. At the lowest level, the parameter $r(t)$ estimates the rate of reward on a specific option. The magnitude of the trial-by-trial change in the reward rate is controlled with a parameter called volatility (v). More specifically, the exponential of v effectively determines the scale of possible updates by setting the width of the transition probability distribution between and $r(t)$ and $r(t+1)$. Finally, the change in volatility is governed or tuned by a parameter k (the second-order volatility, for more details see Behrens et al., (2007)).

Finally, the change-detection Bayesian model is a modification of the model by Jang et al. (Jang et al., 2015) to make it a generative model which can detect changes in reward schedule and choose the better option with a fixed probability. The original model is a post hoc model that tries to predict the subject's choice reversal by estimating the posterior probability that reversals occurred on each trial (Costa et al., 2015). To do so, it assumes that subjects choose the most rewarding option with a certain probability and reverse their choice behavior when odds of occurring a reversal reach a certain threshold. Moreover, the model assumes that subjects update prior about the location of the reversal over time. We made a few modifications to enable this post hoc model to generate choice sequences and measure its performance during the PRL task (Figure 6A). More specifically, we assumed that to estimate the probability that a reversal had occurred on a given trial, the model has access only to reward feedback from the previous trials, and that there is only one reversal across two blocks of trials as in the original model (Costa et al., 2015). Moreover, to obtain the maximum performance for this model, we used a 2D grid search of initial values of prior and the update coefficients of animal's belief about reversal occurrence for each environment (see Jang et al., 2015 for more details).

QUANTIFICATION AND STATISTICAL ANALYSIS

Fitting of experimental data and data analysis

Based on previous results (Donahue and Lee, 2015), we only considered models in which reward probability and magnitude are combined additively, and reward probability values are updated for both targets on every trial (coupled learning). More specifically, the estimated probability of choosing the green target, $P_g(t)$, was fit according the following equation:

$$\text{logit } P_g(t) = \beta_0 + \beta_{stay}D_{pc} + \beta_p\left(p_g(t) - p_r(t)\right) + \beta_m\left(m_g(t) - m_r(t)\right) \quad \text{(Eq. 12)}$$

where β_0 measures an overall bias toward the green or red target, β_{stay} captures the tendency to repeat the previous choice ($D_{pc} = 1$, 0 if the previous choice was green or red, respectively), β_p and β_m are the regression coefficients for reward probability and magnitude, $p_g(t)$ is the estimated probability based on a given model for the green target (equal to $F_{g+}(t)$ and $V_g(t)$ in the RDMP and RL models, respectively), and $m_g(t)$ is the magnitude of the possible reward on the green target. Similarly, $p_r(t)$ and $m_r(t)$ are the estimated probability and the magnitude of the possible reward on the red target, respectively.

We used eight different models to estimate reward probability on each trial in order to fit the experimental data. We utilized the standard maximum likelihood estimation method by minimizing the negative log likelihood to obtain the best fitting parameters for each model. These eight models include: two Bayesian models with different mechanisms for solving the PRL task (hierarchical and change-detection Bayesian); two types of RL models, RL(1) and RL(2), with constant or time-dependent learning rates; the RDMP model with three meta-states; and a simplified version of the RDMP model (see below). To directly estimate the transition probabilities in the RDMP model, we used the architecture presented in Figure 2A with three meta-states ($m = 3$) but allowed any values for these transitions with the constraint that they should be smaller for deeper meta-states.

To assess the goodness-of-fit (negative log likelihood) for each model, we employed 5-fold cross-validation where we fit the data using 80% of randomly selected sessions from a given environment (95 sessions, about 52000 trials) and used the best fitting parameters to predict choice on the remaining 20% of the sessions (23 sessions, about 13000 trials). We repeated this procedure 110 times (i.e., bootstrapped) to obtain a stable average value for the log likelihood for each model. Crucially, the large amount of

behavioral data allowed us to accurately test different models. To compare the results of fits using different models, we only used the test trials. Additionally, to capture the time course of learning rates over time in the sRDMP, RL(1) and RL(2) models (Figures 5A and 5D and Figure S4), we also fit experimental data from each session separately using the maximum likelihood estimation method.

To test the prediction of the RDMP model (time-dependent, choice-specific learning rates) independently of its specific implementation, we used a simplified version of this model referred to as the 'simplified' RDMP (sRDMP). Moreover, using the sRDMP model also allowed us to circumvent degeneracy in the solution (i.e., lack of unique solution) for the RDMP model because each value of the synaptic strength could potentially correspond to many different distributions of synapses in different weak and strong meta-states. Compatibility of fits based on RDMP and sRDMP illustrates that the sRDMP model can be used to detect the contributions of RDMP to behavior. In the sRDMP model, we assumed separate time-dependent learning rates for trials when reward was assigned to the better and worse options. Considering the overall behavior of the effective learning rates over time (Figure 3A) we also assumed the learning rate on trials when the reward was assigned to the better or worse option can exponentially increase or decrease over time after a reversal based on following equation:

$$\alpha(t) = \alpha_{ss} - (\alpha_{ss} - \alpha_0) \times exp^{(-t/\tau)} \quad \text{(Eq. 13)}$$

where α_0 and α_{ss} are the initial and steady-state learning rates, τ is the time constant of the decay, and t is the number of trials after a reversal. According to Equation 13, the learning rates could increase or decrease over time, or stay constant. We used a similar approach for fitting data using the RL(1) and RL(2) models with time-dependent learning rates.

To compare the prediction of the models for choice on congruent and incongruent trials (Figure 4B), we computed the average negative log likelihood (–LL) per trial for each sequence. More specifically, for congruent trials, the average –LL was computed for the last trial in all reward sequences 011, 0111, etc., where 1 and 0 denote reward assignment on the better and worse options, respectively. For incongruent trials, the average –LL was computed for the last trial in all sequences 010, 0110, etc.