Supplemental Information

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

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

Supplemental Data 1, related to Figure 4. Models' response to reward sequence.

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. To examine this prediction and reveal its behavioral consequences, we computed the change in the synaptic strength due to reward assignment as a function of the number (n) of consecutive rewards assigned to the better option in two environments with different levels of volatility. The change in the synaptic strength due to another (congruent) reward assignment on the better option, $\Delta F_C(n)$, was initially smaller for the stable environment but this difference diminished as n increased (Figure S3A). In contrast, when a sequence of rewards assigned to the better option was followed by (incongruent) reward on the alternative option, the corresponding change in the synaptic strength, $\Delta F_I(n)$, was more negative for the volatile environment (Figure S3A).

These results could be understood as follows. A larger fraction of synapses tends to be in unstable weak meta-states in the volatile compared to the stable environment. Therefore, the initial response to the subsequent reward on the better option would be larger in the volatile environment. Similarly, response to subsequent reward on the worse option would be stronger (more negative) in the volatile environment because larger fractions of synapses occupy unstable strong meta-states. As more rewards are repeatedly assigned to the better option, most synapses associated with that option transition to stable strong meta-states. Therefore, eventually, the response in both environments goes to zero for large n since only a small fraction of synapses would be in weak meta-states.

Overall, these results indicate that the model's response to reward on the better option following consecutive reward assignment on that option ('congruent' sequence) is larger for the volatile than the stable environment. Similarly, the model's response to 'incongruent' reward assignment (i.e. reward on the worse option following consecutive reward assignment on the better option) was stronger (more negative) for the volatile environment (Figure S3A). Because the synaptic strength determines the model's choice behavior, the aforementioned pattern of changes in the synaptic strength is also reflected in changes in choice probability. More specifically, the first few consecutive congruent outcomes caused a larger increase in choice preference for the better

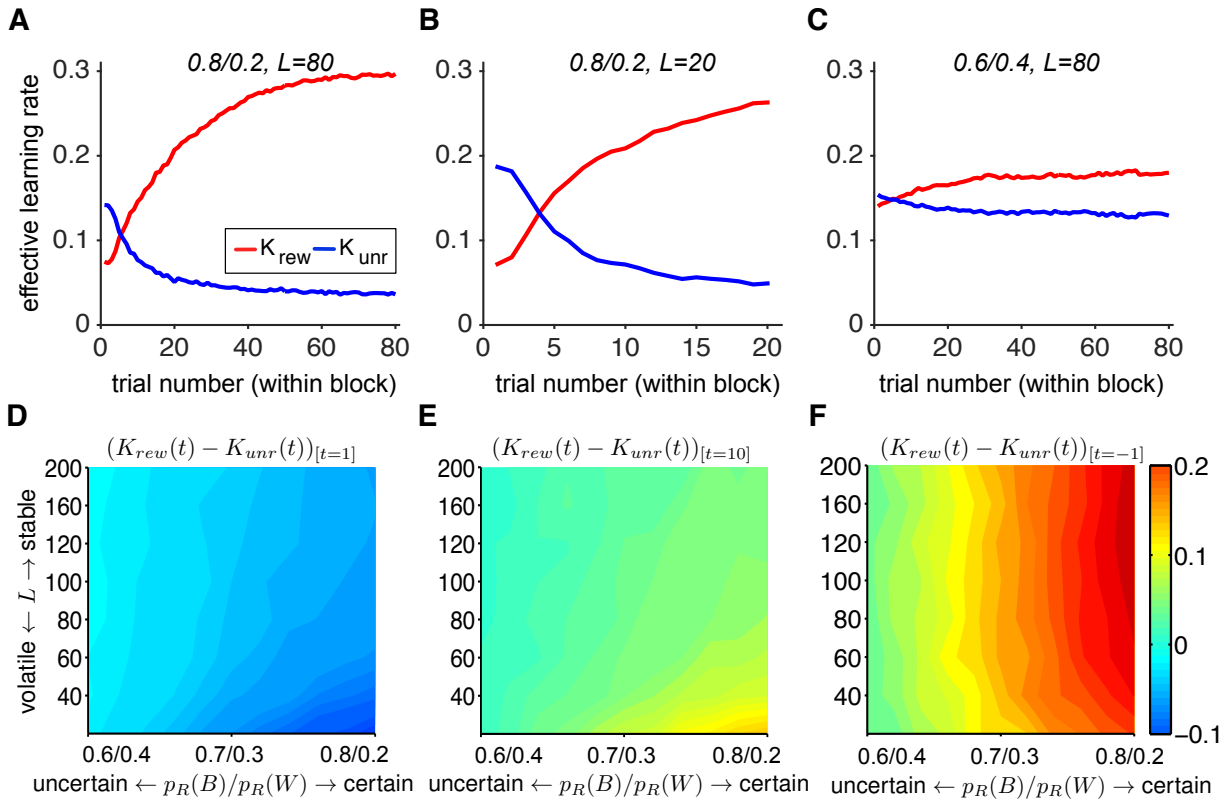
option in the volatile than in the stable environment, but this change became similar for longer sequences (Figure S3C). In addition, the preference for the better option was reduced after an incongruent outcome more in the volatile than in the stable environment, and this difference between the two environments was relatively preserved even for longer sequences (Figure S3F).

To test whether this pattern of response is unique to our model, we next measured the response to reward feedback in the RL(2) and hierarchical Bayesian models. This analysis could not be applied for the change-detection Bayesian model, since this model selects the better option with a fixed probability. For the RL(2) model, changes in the value function for both congruent and incongruent trials, $\Delta V_C(n)$ and $\Delta V_I(n)$, were initially larger in the volatile than the stable environment, but this difference disappeared with longer sequences of reward assignment (Figure S3B). This happens because changes in the value function in this model are determined by reward prediction error, which is on average larger (smaller) for the volatile than the stable environment for congruent (incongruent) trials. The difference between the two environments disappears as reward prediction error approaches 0 and -1 for longer congruent and incongruent sequences, respectively. The change in choice probability due to congruent and incongruent reward feedback followed the same pattern as the $\Delta V_C(n)$ and $\Delta V_I(n)$ (Figure S3D,G).

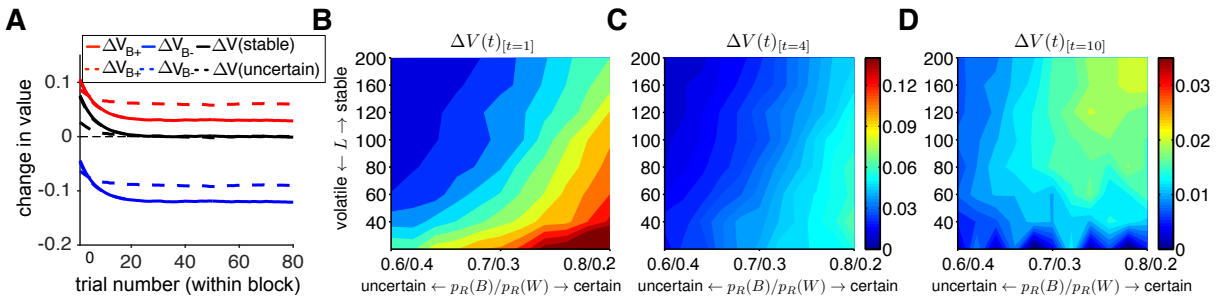
The hierarchical Bayesian model showed a different pattern of response to reward feedback. For the shortest congruent sequence, the Bayesian model changed its behavior more strongly in the volatile than in the stable environment, but this response to consecutive reward assignment diminished more quickly in the volatile environment as n increased (Figure S3E). By contrast, following incongruent sequences of reward assignment, the preference for the better option was reduced more in the volatile environment than in the stable environment and this difference between the two environments was preserved even for longer sequences (Figure S3H). The hierarchical Bayesian model responds this way because it considers the volatility of the environment for integration of reward outcomes. Overall, these results demonstrate that our model based on RDMP provides specific behavioral predictions which are qualitatively different from the predictions of the RL(2) and hierarchical Bayesian models.

We note that it is not feasible to directly test which predicted patterns of response to congruent and incongruent trials is more compatible with experimental data, since those patterns were

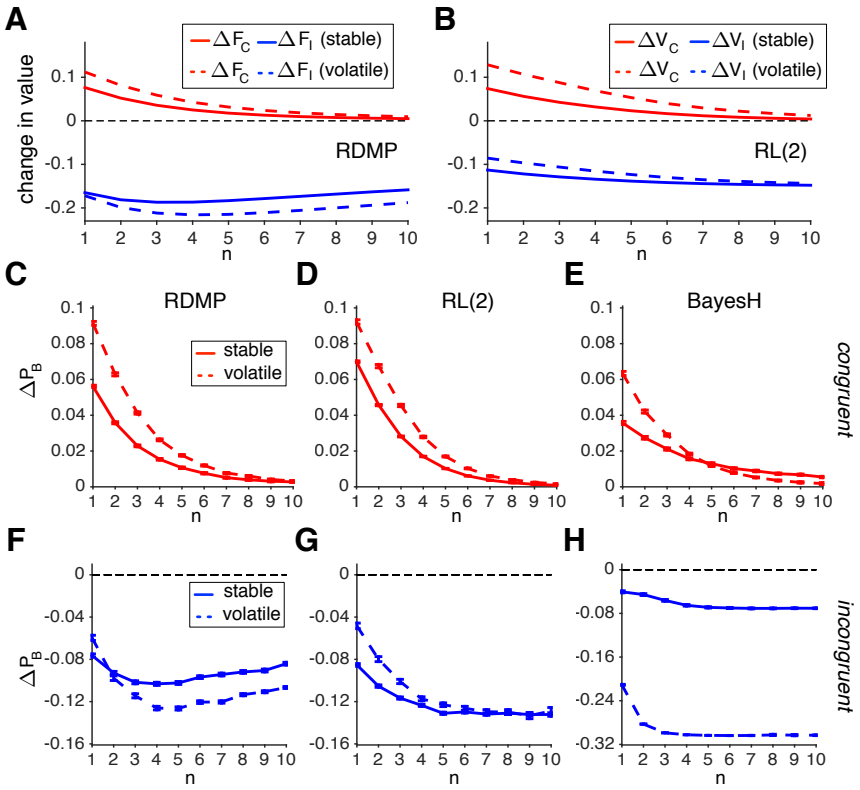
revealed using a large number of simulated data. Therefore, we tested these predictions using the fit of monkeys' choice on congruent and incongruent trials.



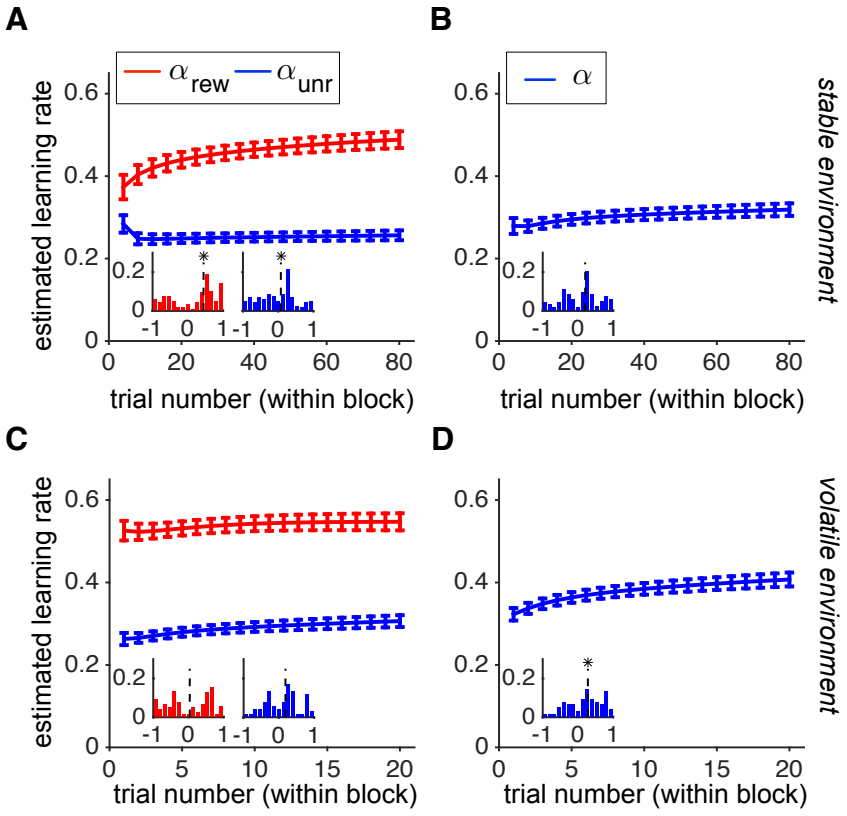
Supplemental Figure 1, **related to Figure 3.** The RDMP model shows a larger effective learning rate on rewarded than unrewarded trials. (**A-C**) Plotted are the effective learning rates for rewarded (K_{rew}) and unrewarded (K_{unr}) trials as a function of the trial number after a reversal in three different environments. With enough reward feedback, the effective learning rate on rewarded trials becomes larger than the effective learning rate on unrewarded trials. (**D-F**) The difference in the effective learning rates on rewarded and unrewarded trials at three time points after a reversal in different environments. The difference is more negative for more certain and/or volatile environments right after reversals, but becomes positive at the steady state (t = -1). At the steady state, the difference between the effective learning rates on rewarded and unrewarded trials mainly decreases with the uncertainty in the environment.



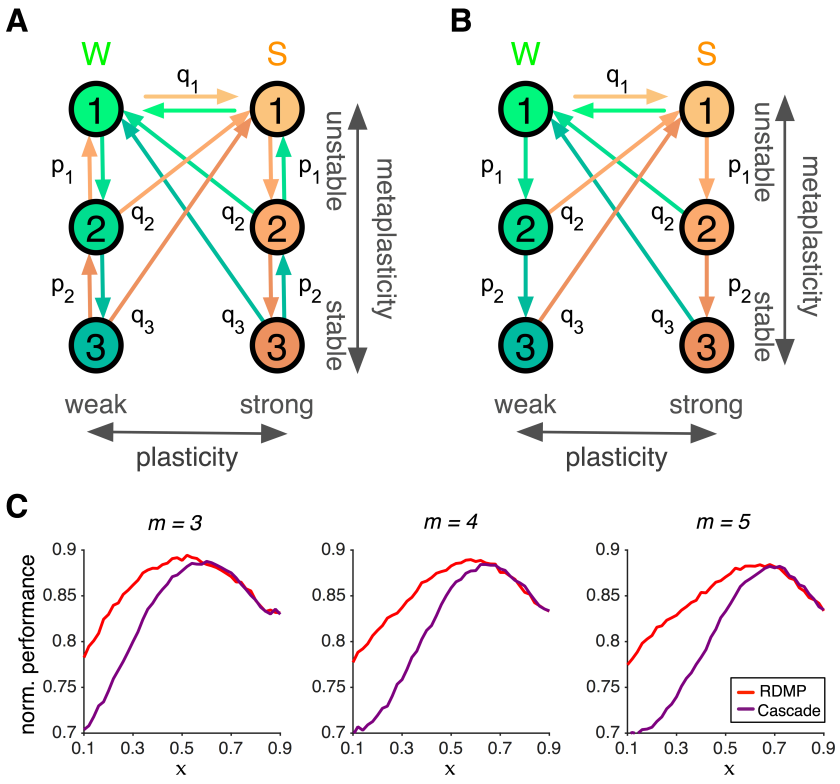
Supplemental Figure 2, related to Figure 3. Changes in the response of the RL(1) model to reward feedback over time. (**A**) Plotted are the change in the value function in response to reward assignment on the better (ΔV_{B+}) or worse (ΔV_{B-}) option, as well the overall change in the value function (ΔV) as a function of the trial number after a reversal in the stable (solid) and uncertain (dashed) environments. For these simulations, the learning rate was set to 0.15. (**B-D**) The overall change in the value functions at three time points after a reversal and in different environments, using the optimal learning rate in each environment. The model's response to reward feedback was stronger for more certain and/or volatile environments right after reversals and this difference decreased over time.



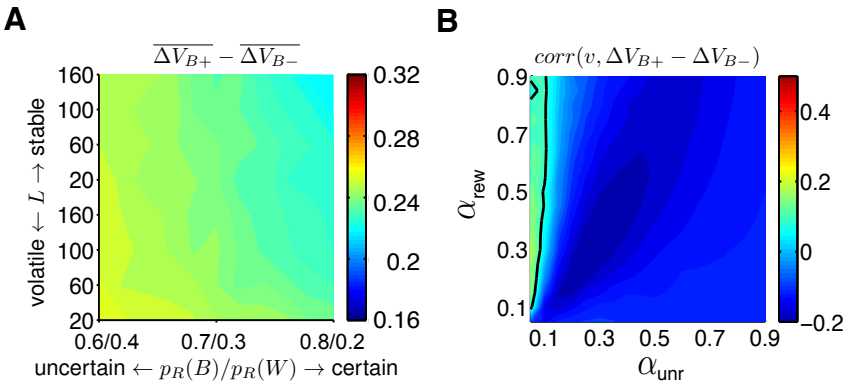
Supplemental Figure 3, related to Figure 4. Adjustment of learning to specific sequences of reward assignment in different models. (**A**) Predictions of the RDMP model. Plotted in red is the change in the synaptic strength due to reward assignment on the better option (ΔF_C) as a function of the number (n) of preceding consecutive rewards assigned to that option in two environments with different levels of volatility. Reward sequence 011, where 1(0) denotes reward on the better (worse) option, corresponds to a congruent sequence with n = 1, etc. Blue curves show the change in the synaptic strength when a sequence of reward assignment on the better option was followed by reward on the alternative option (ΔF_I). (**B**) The same as in **A** but predicted by the RL(2) model. (**C-E**) Predictions of three different models for the change in probability of choosing the better option on congruent trials as a function of the number of consecutive rewards assigned to the better option (BayesH: hierarchical Bayesian model). The error bars indicate s.e.m. which are smaller than the marker in many cases. (**F-H**) The same as in **c-e** but for incongruent trials. Reward sequence 010 corresponds to an incongruent sequence with n = 1, etc. Overall, the three models showed qualitatively different patterns or response to congruent and incongruent sequences of reward assignment in the two environments.



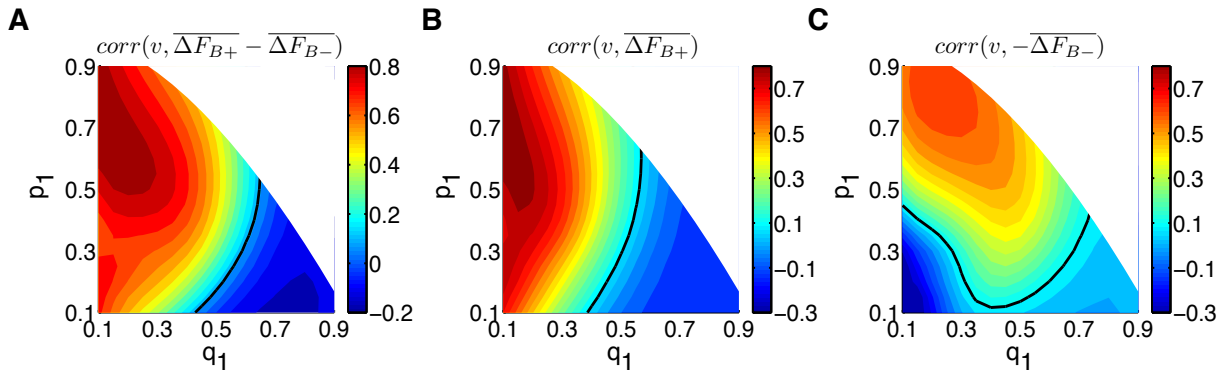
Supplemental Figure 4, related to Figure 5. The estimated learning rates over time based on the session-by-session fit of monkeys' choice behavior during the modified PRL task, using the RL models with time-dependent learning rates. (**A**) The estimated learning rates for rewarded and unrewarded trials using the RL(2) model in the stable environment. 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 ($p < .05$) different from zero. (**B**) The estimated learning rate using the RL(1) model in the stable environment. (**C-D**) The same as **A-B** but for behavior during the volatile environment.



Supplemental Figure 5, related to Figure 6. Comparison of the RDMP and the cascade model. The RDMP model outperforms the cascade model of metaplasticity because of the ability to destabilize weak (strong) synapses while stabilizing strong (weak) synapses on potentiation (depression) events. (**A-B**) Architecture of the RDMP (**A**) and cascade model (**B**). To use the same set of parameters in two models we defined all transition probabilities based on a single parameter x (as in Fusi et al. (2005)). More specifically, we set $p_i^+ = (x)^i,\ p_i^- = (x)^i$ for $1 \leq$ i $< m$ and $q_i^+ = (x)^i,\ q_i^- = (x)^i$ for $1 \leq$ i $\leq$ m . (**c**) Plotted is the normalized performance of the RDMP and cascade models as a function of the single transition probability x and with different number of meta-states, in a universe with many different levels of uncertainty/volatility. The normalized performance of 0.7 corresponds to chance performance. This happens due to the probabilistic nature of PRL task where even the performance of the omniscient observer cannot surpass $p_R(B)$ in a given environment while choosing randomly results in performance of 0.5. The RDMP model outperforms the cascade model for most of value of x, and the difference in performance between the two models increases with a larger number of meta-states.



Supplemental Figure 6, related to Figure 7. Lack of neural correlates of estimated volatility in the RL(2) model. (**A**) The average value of the difference in the value functions in the RL(2) model in different environments. There is only a small modulation of the difference in value functions by uncertainty and volatility. (**B**) The correlation coefficient between trial-by-trial estimate of $(\Delta V_{B+}(t) - \Delta V_{B-}(t))$ and estimated volatility by the hierarchical Bayesian model over a wide range of model's parameters during ten environments with different levels of volatility. Area to the left of the black curve indicates parameter values for which the correlation is larger than 0.1. For easier comparison, the ranges of values in the color-maps are set similar to those in Figure 7.



Supplemental Figure 7, related to Figure 7. The Correlations between the change in synaptic strengths in the RDMP model and estimated volatility by the hierarchical Bayesian model depends on whether reward is assigned to the better or worse option. Plotted is the correlation between the average time course of estimated volatility by the hierarchical Bayesian model and the difference between changes in the synaptic strengths $(\Delta F_{B+}(t) - \Delta F_{B-}(t))$ (**A**), changes in the synaptic strength when reward is assigned to the better option $\left(\Delta F_{B+}(t)\right)$ (**B**), and the absolute changes in the synaptic strength when reward is assigned to the worse option $(-\Delta F_{B-}(t))$ (**C**). The black curves indicate correlation values equal to 0.1. Note that we obtain larger correlation values in **A** than in Figure 7C, because we computed the correlation using the average quantities over blocks of trials in the former instead of the actual quantities on each trial in the latter. These correlations are measured over a wide range of model's parameters (the maximum transition probabilities) and during ten environments with different levels of volatility. Overall, the correlation was larger for when reward is assigned to the better option than when reward was assigned to the worse option.