Neuron, volume *66*

Supplemental Information

States versus Rewards: Dissociable Neural Prediction Error Signals Underlying Model-Based and Model-Free Reinforcement Learning

Jan Gläscher, Nathaniel Daw, Peter Dayan, and John P. O'Doherty

Supplemental Figure S1

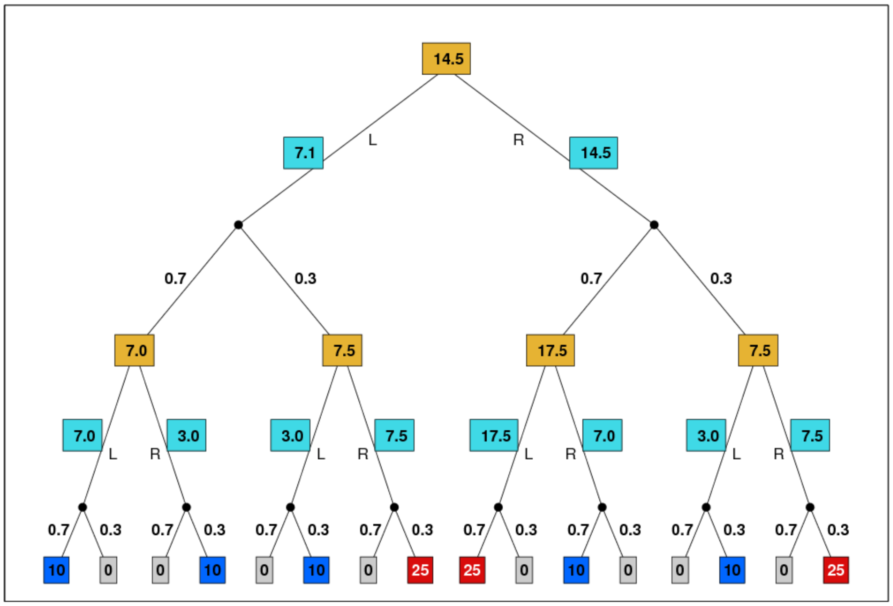

Figure S1, related to Figure 1. Optimal state-action *Q(s,a)* (cyan) were computed by multiplying reward magnitude and transition probabilities along each path through the decision tree. State values *Q(s)* (orange) maximize the state-action values available in that state. These optimal *Q*-values were used to define a correct choice in each state, which was correlated with the state prediction error signal in the non-rewarded session 1.

Supplemental Figure S2

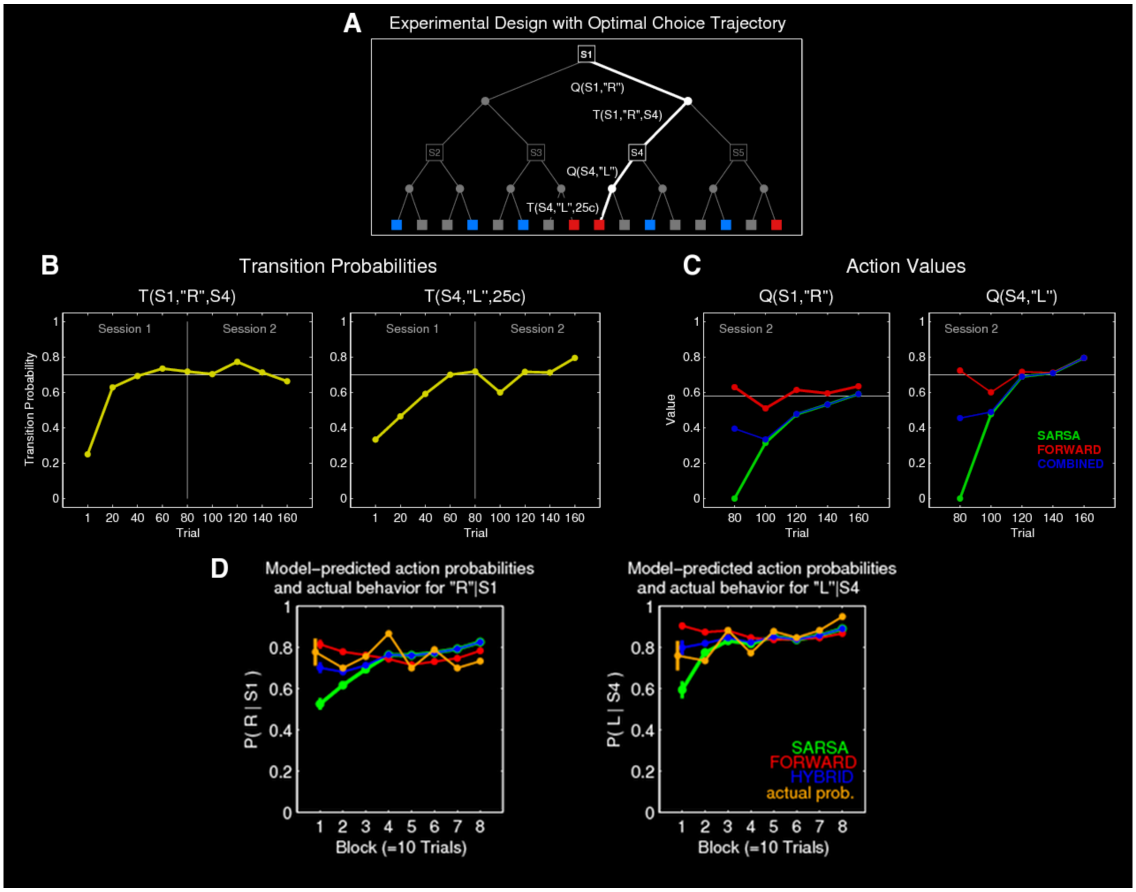

Figure S2, related to Figure 2. Visualization of the evolution of the transition probabilities and action values, and action probabilities for the optimal choice trajectory during the course of the experiment, averaged across the simulation of each subject. **(A)** The optimal path through the decision tree. **(B)** The convergence of the estimated transition probability (yellow) toward the optimal probability of 0.7 indicates that toward the end of the first scanning session, the FORWARD learner has successfully acquired this part of the state space. **(C)** Action values are plotted as the difference between the optimal and non-optimal action in each of the two states along the optimal trajectory. Only the FORWARD learner distinguishes these two options at the beginning of Session 2; the SARSA learner assigns no value to either. With subsequent experience, the SARSA learner also learns to assign a higher value to the better choice. Whereas the FORWARD value predictions reflect the true difference in the actions' values, the SARSA model slightly overestimates the difference due to having had limited experience with the consequences of the poorer choice. The blue curve shows the net value difference for the HYBRID learner after taking the weighted combination into account. Because of the rapid decline of the weight of the FORWARD learner the combined value difference tracks that of the SARSA learner for the balance of session 2. That is, in the model's fit, behavior is initially determined by the FORWARD learner, but the progress of subsequent updates is quickly dominated by SARSA learning. **(D)** The model-prediction action probabilites in session 2 for the optimal action in both states show a similar pattern as the action values in (C). In addition, the actual action probabilities derived from the behavioral choice data are shown in orange. To avoid visual cluttering of the display, errorbars (s.e.m. across subjects) are only shown at the first data point. Even by visual inspection the close correspondences between model predictions and data indicate that the computational models employed here fit the data well.

Supplemental Figure S3

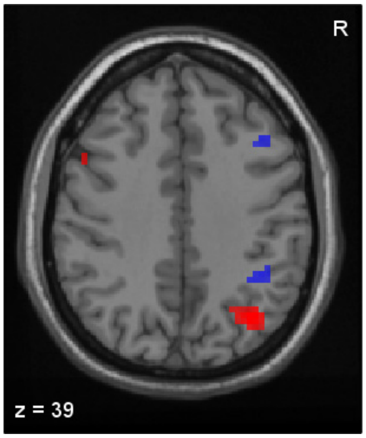

Figure S3, related to Figure 3. Distinct anatomical locations in the parietal cortex for state prediction error (red) and unsigned reward prediction error (blue). The SPE (conjunction from both sessions) correlates with the BOLD activation in the posterior IPS and angular gyrus, whereas the unsigned RPE correlates with the anterior IPS (both contrasts thresholded at $p < 0.001$ uncorrected). Both error signals were entered unorthogonalized into the SPM design matrix

Supplemental Figure S4

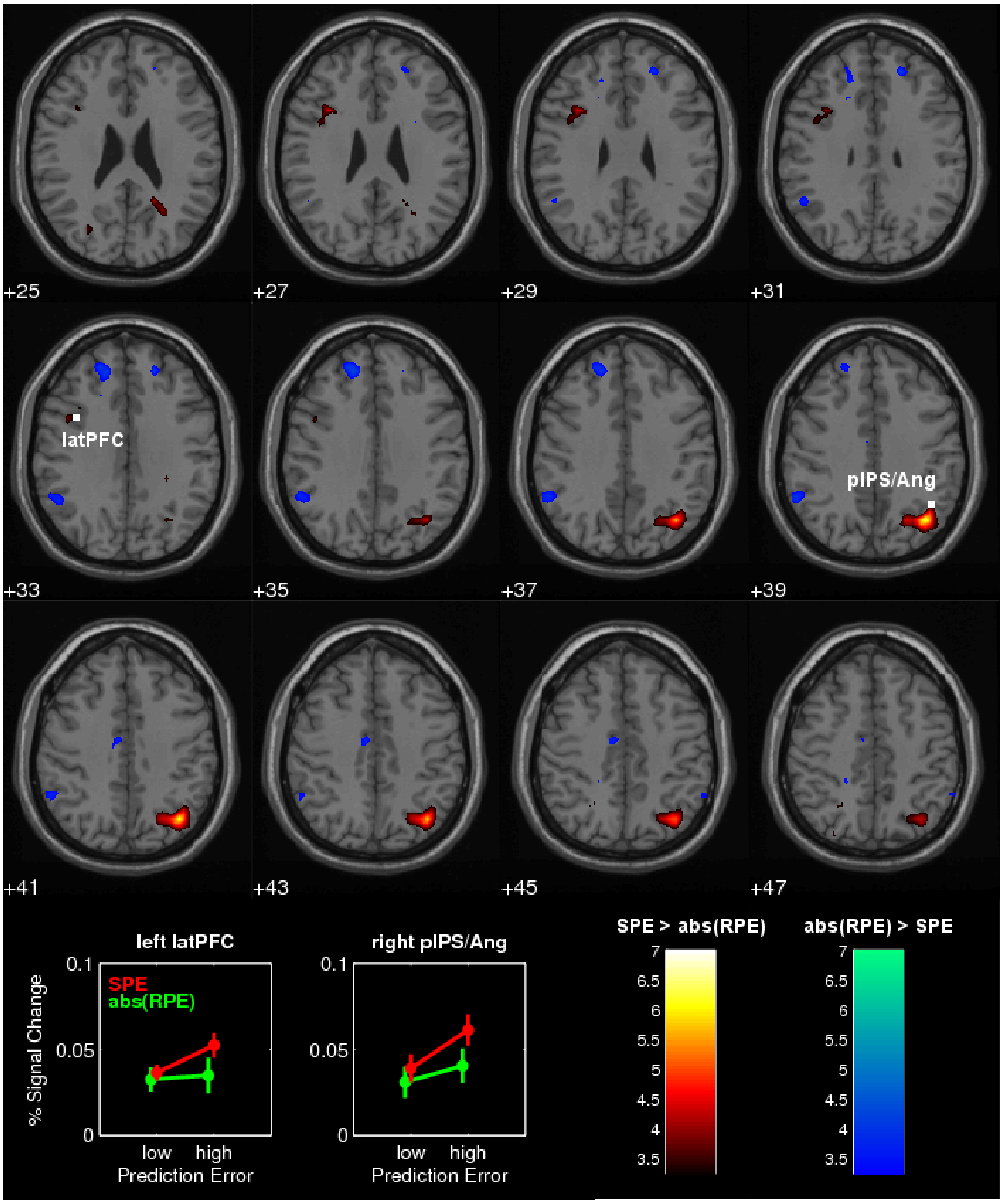

Figure S4, related to Figure 4. Differential contrast comparing the average SPE (from both sessions) and the absolute value of the RPE (abs(RPE)). Both error signals were entered into the same design matrix. SPMs are shown thresholded at $p < 0.001$ uncorrected. This contrast reveals that activity in right pIPS is significantly better explained by the SPE signal than by the abs(RPE) signal (at $p < 0.05$ corrected), while activity in left latPFC shows a similar effect (at $p < 0.001$ uncorrected). The graphs below show the average percent signal change (across subjects) for the trials with low and high SPE and abs(RPE) (median split) in the two target regions (right pIPS/angular gyrus and left lateral PFC). These plots were extracted from a 10 mm sphere centered on the peak coordinates in pIPS and lateral PFC (white dots in the figure above) from the conjunction analysis between the SPE signals in session 1 and session 2 (see Figure 4 in the main text and Table 2). Taken together these results suggest that activity in the regions identified as correlating with SPE do not correlate with abs(RPE), arguing against the possibility that the SPE signal reflects a non-specific arousal signal.

Determining the weighting function in the HYBRID learner

The HYBRID learner defines a weighting function that negotiates between the model-free SARSA learner and the model-based FORWARD learner in the second free-choice session (see Methods for model equations). We tested a constant, linear, and exponential weighting function and chose the latter, because it provided the best model fit (as determined by the negative model likelihood, see below).

Table S1, related to Figure 2. Model parameters, negative model likelihoods, and Akaike Information Criteria (AICs) for a constant, linear, and an exponential HYBRID learner.

Parameter	exponential weighting function	constant weighting function	linear weighting function
SARSA learning rate	0.20	0.40	0.34
FORWARD learning rate	0.21	0.20	0.21
Offset for exp. decay	0.63		
Slope of exp. decay	0.09		
Constant weight[1]		0.21	
Intercept for linear function[2]			0.33
Slope for linear function			-0.0042
Inverse softmax temperature	4.91	5.18	5.07
Number of model parameters	5	4	5
Negative model likelihood	1202.28	1207.28	1204.49
AIC	2414.56	2422.56	2418.98

[1] influence of model-based FORWARD learner

[2] initial influence of model-based FORWARD learner at beginning of session 2

Supplemental Table S2

Table S1, related to Figure 5. Probability of correctly predicted choices and pseudo-R^2 for each subject based on the fit of the HYBRID model. Pseudo-R^2 are computed as (R – L) / R (Daw et al., 2006) for each subject, where L and R are the negative log likelihoods of the HYBRID model and a null model of random choices respectively. Notice that the pseudo-R^2 measures are low in some participants because the individual likelihoods are derived from the single set of parameters fitted across the entire sample, which is used throughout this paper.

Subject	**Pseudo-R^2**	**P(corr. pred. choice)**
1	0.03	0.69
2	0.28	0.79
3	0.16	0.73
4	0.52	0.84
5	0.05	0.68
6	0.80	0.98
7	0.79	0.97
8	0.22	0.74
9	0.63	0.91
10	0.01	0.64
11	0.05	0.66
12	0.54	0.89
13	0.81	0.98
14	0.48	0.82
15	0.35	0.81
16	0.70	0.91
17	0.47	0.89
18	0.32	0.74
mean	**0.40**	**0.81**

Supplemental Experimental Procedures

Reward exposure between scanning sessions

After completing the non-rewarded fixed-choice scanning session 1, in which our participants acquire knowledge about the state transition probabilities, they were confronted with the rewards that they would earn at each of the 3 outcomes states in the subsequent free-choice scanning session (see Figure 1b for the display that the participant observed).

Prior to the second scanning session this mapping of outcome state to rewards was rehearsed by all of our participants in a simple choice task. They either saw one or two of the three outcome states. If they saw only a single outcome state, they were instructed to press the button that corresponded to the side of the screen that the outcome state was presented in. If they observed two outcome states, they would have to pick the one that gave them the higher payoff by pressing the corresponding button. After their button press, they saw the string "You won X cents.", where X was replaced by the corresponding outcome.

All outcome states were systematically paired with each other and presented 4 times with randomized positions on the screen (left or right). All single outcome states were also presented 4 times totaling 24 trials in this reward mapping rehearsal task. The outcomes of each trial were added to the participant's total payoff.

Performance on this reward mapping rehearsal was very high. Of our 18 participants only 3 made any errors at all (2 participants missed on 2 trials, 1 participant on 3 trials) resulting in 98.8% correct performance across participants.

Creating a random null distribution for testing the effects of model-based learning

Under model-based learning, exposure (even if guided as in session 1) leads to the build-up of a specific state space representation in each subject that can be utilized for making subsequent choices. However, if the subjects did not learn anything about the state transitions in session 1, then a random trial sequence in a randomly permuted state space should lead to the same qualitative model fit of the HYBRID learner than the original trial sequence with the actually experienced state transition. To further support the evidence for model-based learning, we use this reasoning to create a null distribution of model likelihoods under random state transitions and trial sequences (using 1000 bootstrap samples) against which the original model fit can be compared.

For each bootstrap sample we randomly permuted the trial sequence in session and the position of the intermediate states (2nd layer in the decision tree) for each subject, while keeping the session 2 data (free choices) intact. We then refitted the HYBRID model under these random conditions and recorded the final model likelhood. This likelihood distribution represents the null hypothesis of no model-based learning against which the model fit of the original data was compared. This analysis yielded a highly signficant effect in favor of model-based learning: 99.6% of the likelihoods of the permutation samples were worse than the original model likelihood ($p = 0.004$).