Supplementary Information Appendix for

How Instructed Knowledge Modulates the Neural Systems of Reward Learning

Jian Li, Mauricio R Delgado, and Elizabeth A Phelps

Supplementary Methods

Experimental procedures. On each trial, participants were instructed to select the visual cue if they believed the number underneath the cue was greater than 5, and to select the number 5 if they believed the number underneath the cue was less than 5. The underlying number was then revealed. If the participant was correct, s/he was rewarded with $1.00 and if the participant was incorrect s/he was punished by taking away $0.50. The participant received the sum of monetary outcomes from a randomly selected set of 30 trials across both sessions at the end of the experiment in addition to the standard compensation. Prior to the instructed session, a thorough description of the specific cues linked to different probabilities ($P \in \{25\%, 50\%, 75\%, 100\%\}$) was presented to participants and they were required to memorize these cue-probability associations successfully before the session started. In addition, the probabilities were displayed on top of each cue during the instructed session to remove any uncertainty as to whether the participants were aware of the correct cue-probability association. In the feedback session, there was no instruction of the probabilities linked to specific cues prior to the session and these probabilities were not displayed, so that the only means of learning the probabilities was trial and error feedback.

fMRI image acquisition. Participants laid supine with their heads in the scanner and observed the rear-projected computer screen via a mirror mounted to the

head coil. Visual stimuli were presented using Psychtoolbox (http://www.psychtoolbox.org). Choices made by participants were registered using two MRI-compatible button boxes. Thirty-nine contiguous axial slices (3 x 3 x 3 mm voxels) parallel to the AC-PC line were obtained while participants were engaged in both sessions of the task. High-resolution T1-weighted scans were acquired in the same location as the functional images. The functional run was echo-planar imaging (EPI) with gradient recalled echo with a repetition time (TR) was 2000 ms. The echo time (TE) was 25 ms with a flip angle of 90^{o} and 64 x 64 within plane resolution. The data were then analyzed using SPM2 (Wellcome Department of Cognitive Neurology, London, UK) and xjView (http://www.alivelearn.net/xjview/). All images were slice-timing corrected. Participants' head movements were estimated using a rigid-body translation with 6 parameters (x, y, z, pitch, roll and yaw) and all 20 participants were confirmed with head movements less than 2 mm in each direction. Individual T1 images were segmented to obtain the individual gray matter images that were later normalized to the MNI template to obtain the transformation matrix. The transformation matrix was then applied to all the functional images to normalize them into MNI space. Images were then smoothed using a 6 mm isotropic Gaussian kernel and high-pass filtered (128s) in the temporal domain.

Model Fitting and selection. We fitted two models to subjects' behavior data in the feedback session and three models in the instructed session to compare the performance of most influential models in recent literature (1, 2).

Feedback Session:

Q-Learning with single learning rate. Subjects were not given any information about the reward probability of each stimulus. We assume the initial Q-values for the stimuli were the same ($Q_{25\%} = Q_{50\%} = Q_{75\%} = Q_{100\%}$). The update of each stimulus action value is based on the following rules:

$$\delta(t) = r(t+1) - Q_c(t)$$

$$Q_c(t+1) = Q_c(t) + \alpha\delta(t)$$

Where $\delta(t)$ is the prediction error at time t, r is the obtained reward (normalized between [0 1]), $c \in$ {No.5, $cue_{25\%}$, $cue_{50\%}$, $cue_{75\%}$, $cue_{100\%}$} and α is the learning rate.

Q-Learning with different learning rates for positive and negative PEs. Stimuli action value is updated according to the following rules:

$$\delta(t) = r(t+1) - Q_c(t)$$

$$Q_c(t+1) = Q_c(t) + \alpha_G\delta_+(t)$$
$$Q_c(t+1) = Q_c(t) + \alpha_L\delta_-(t)$$

where α_G and α_L are learning rates associated with positive and negative (δ_+ and δ_-) prediction errors.

Free parameters in the models were estimated using the log likelihood estimate (i.e. $\log(\prod_t P_*(t))$). Performance of both models was compared using the Bayesian information criterion (BIC):

$$BIC = -2 \cdot \ln(L) + k \cdot \ln(n)$$

where L is $\prod_t P_*(t)$, *k* is the number of free parameters and *n* is the number of trials (see Table S1). RL model with single learning rate tended to perform better in the feedback session and the PE regressor was sequentially generated from this model for neuroimaging linear regression analysis.

Instructed Session:

Q-Learning with single learning rate. In this session, subjects were provided extra instruction about the reward probability for each visual stimulus. It's thus reasonable to believe that the initial Q-values ($Q_{25\%}$, $Q_{50\%}$, $Q_{75\%}$, $Q_{100\%}$) for the stimuli were different. The update of each stimulus action value is the same as in the feedback session.

Q-Learning with different learning rates for positive and negative PEs. The action value updating approach is the same as in the feedback session.

Q-Learning with "confirmation bias". This model proposes that the instructed stimulus activate the striatal Go representations and increases the effect of positive prediction error (δ_+) following the instructed choice, while also diminishing the effect of negative prediction error (δ_-) when the instructed choices receive punishing feedback (1). This "confirmation bias" is implemented as following:

$$\delta(t) = r(t+1) - Q_i(t)$$

$$Q_i(t+1) = Q_i(t) + \alpha_I \alpha_G \delta_+ + \frac{\alpha_L}{\alpha_I} \delta_-$$

where α_G and α_L are learning rates associated with positive and negative (δ_+ and δ_-) prediction errors; α_I ($1 \leq \alpha_I \leq 10$) that amplifies gains and reduces losses following the instruction.

The above three models were compared using BIC and the RL model with different learning rates for δ_+ and δ_- tended to be the marginally best model (Table S2).

The quality of the model fitting is quantified by how well they are able to account for the actual pattern of participants' choice. The log likelihood estimate (i.e. $\log(\prod_t P_*(t))$) was maximized to determine the free parameters in the models. To avoid local minima in parameter fitting, 30 randomly selected starting points were initiated and the best-fit values were taken across all final parameters values. The following restrictive rules also applied: $0 \leq \alpha \leq 1$; $0 \leq \alpha_+ \leq 1$; $0 \leq \alpha_- \leq 1$; $0 \leq m \leq 50$ and $0 \leq Q_c \leq 1$.

To make the results more comparable between the feedback and instructed sessions, we adopted the simplest RL model by varying only the learning rate and the slope of the softmax choice function of the *Q-learning* model to fit subjects' behavior in both sessions using the same maximum likelihood algorithm described above. Initial Q values were set at 0 in both sessions. We extracted the prediction error term from the best-fitting-parameter simple RL model for both the feedback and instructed sessions and examined the neural correlates of PEs in both sessions (Fig. S3). The results were similar to what we reported in the main text (Fig. 3).

Behavioral analysis:

In the feedback session, since participants have no reason to have a differential preference towards different visual cues, we assume individual participants start the session with the same internal estimate of the action value (***Q***) expected from each cue (i.e. initial cue value of four different probabilities and the fixed number $\boldsymbol{Q}_{25\%} = \boldsymbol{Q}_{50\%} = \boldsymbol{Q}_{75\%} = \boldsymbol{Q}_{100\%} = \boldsymbol{Q}_{\text{No.5}}$). In particular, we assume the action value (***Q***) was updated according to a Rescorla-Wagner learning rule. However, since participants were instructed by the experimenter about the probability of the action value for each visual cue in the instructed session, the initial weights of four different cues and the fixed number ($\boldsymbol{Q}_{25\%}$, $\boldsymbol{Q}_{50\%}$, $\boldsymbol{Q}_{75\%}$, $\boldsymbol{Q}_{100\%,}$ $\boldsymbol{Q}_{\text{No.5}}$) were treated as different values and their exact values were determined by the best model fitting in the instructed session (see supplementary materials for details).

For both sessions, the probability of choosing a given action is predicted by the model according to a sigmoid function with slope *m*:

$P_{No.5}(t) = \frac{e^{mQ_{No.5}(t)}}{e^{mWQ_{No.5}(t)} + e^{mQ_{cue(i)}(t)}}$, where *cue(i)* ∈ {25%, 50%, 75%, 100%}. In both sessions, for each choice (denote the choice by *c* and *c* ∈ {No.5, $cue_{25\%}$, $cue_{50\%,}$ $cue_{75\%,}$ $cue_{100\%}$}), the reward experienced by the participant *r(t)* was compared with the current modeled weights $Q_c(t)$ to produce the prediction error $\delta(t)$: $\delta(t) = r(t+1) - Q_c(t)$. The prediction error signal was then served as a learning signal to update modeled action values by the amount governed by learning rate α: $Q_c(t+1) = Q_c(t) + \alpha\delta(t)$.

PPI analysis:

We first identified a seed region that showed increased activation during the outcome phase when instructed knowledge was available. The time series of this region was deconvolved based on the assumption that the BOLD signal is the convolution product of underlying neural activity and a canonical hemodymanic response to obtain the time series of underlying BOLD activity. A new general linear model (GLM) was then constructed with the following regressors: 1) Interaction between the BOLD activity in the seed region and a dummy indicator for positive or negative outcomes. 2) The indicator function for positive and negative outcomes. 3). The original BOLD time series in the seed area. The first two regressors were also convolved with a canonical form of the HRF so that observed BOLD signal would be a linear combination of these 3 regressors. We were interested in the neural correlates of the first regressor (Fig. 5*B*), since it identifies the brain areas whose activities showed an outcome specific neural connectivity with the seed area. More specifically, we wanted to identify areas in which the correlation in BOLD activity increases (more negative correlation) during positive outcome trials (wins), given that the majority of our trials yielded positive outcomes.

The BOLD responses reported in the paper are whole brain corrected ($p < 0.05$) at the cluster level (family wise error; FWE) based on the random field theory (RFT) and through a SPM2 plug-in implementation. (3, 4)

Supplementary Tables

Table S1. Free parameters and quality of behavioral fits to 1600 choices from 20 subjects in feedback session. –LL: log likelihood; Pseudo-R^2: McFadden's pseudo R-square; BIC: Bayesian information criterion.

Instructed Session			
	Simple Q	**Q (different rates for gain and loss)**	**Doll's Q (PFC-BG)**
$Q_{no.5}$	0.16	0.26	0.26
$Q_{25\%}$	0.00	0.00	0.00
$Q_{50\%}$	0.19	0.32	0.32
$Q_{75\%}$	0.43	0.75	0.75
$Q_{100\%}$	0.56	1.00	1.00
α_G	-	0.05	0.05
α_L	-	0.00	0.00
α_I	-	-	1.00
α	0.02	-	-
m	10.93	6.25	6.26
LLE	-433.66	-429.06	-429.06
Pseudo-R^2	0.61	0.61	0.61
parameter	7	8	9
BIC	918.96	917.14	924.52
AIC	881.32	874.12	876.12

Table S2. Free parameters and quality of behavioral fits to 1600 choices from 20 subjects in instructed session. –LLE: log likelihood; Pseudo-R^2: McFadden's pseudo R-square; BIC: Bayesian information criterion.

Feedback Session		
	Simple Q	**Q (different rates for gain and loss)**
$Q_{no.5}$	0.00	0.00
Q_{Risky}	0.02	0.04
α_G	-	0.33
α_L	-	0.19
α	0.24	-
m	4.38	3.76
LLE	-550.21	-549.17
Pseudo-R^2	0.50	0.50
parameter	3	4
BIC	1122.55	1127.85

Table S3. Maximally activated voxels in areas exhibiting significant correlation with prediction error (PE) signals in feedback session.

Region	*L/R*	*BA*	*MNI Coordinates*	*Z*
Inferior Parietal Lobe	R	40	51 -39 54	4.60*
Inferior Parietal Lobe	L	40	-36 -60 45	3.98*
Superior Frontal Gyrus	R	10	27 63 -6	4.45*
Middle Frontal Gyrus	R	9	42 42 33	3.73*
Thalamus	R	-	12 -15 12	3.90*
Putamen	L	-	-27 3 0	3.59*
Posterior Cingulate	R	23	3 -33 24	3.43
Precuneus	-	7	0 -75 48	3.38*
Middle Frontal Gyrus	L	10	-42 42 -12	3.24*
Midbrain	R	-	3 -21 -24	3.24
Medial Frontal Gyrus	L	8	-3 30 45	3.08

*significant at $p < .05$ after whole brain cluster correction with a t threshold of 2.54 and an extent of 91 voxels.

For completeness, peaks are reported for all clusters ≥ 15 voxels at $p < .005$ unc.

Table S4. Maximally activated voxels revealed by the win-loss contrast across both the feedback and instructed sessions.

Region	*L/R*	*BA*	*MNI Coordinates*	*Z*
Ventral Striatum	L	-	-3 12 -9	6.12*
vmPFC	L	32	-6 51 -6	5.94*
Angular Gyrus	L	19	-42 -72 30	5.55*
Middle Frontal Cortex	L	9	-21 33 42	4.99*
Superior Temporal Gyrus	R	22	63 -6 0	4.64*
Hippocampus	L	28	-18 -12 -21	4.31*
	R	35	27 -27 -24	3.67*
Fusiform	L	37	-27 -48 -15	4.26
Posterior Cingulate	L	23	-9 -57 12	5.04*
	R	23	12 -54 12	4.28
Superior Temporal Gyrus	L	21	-63 -21 6	3.95*
Temporal Gyrus	R	41	42 -30 9	3.70
Occipital Lobe	L	18	-12 -93 18	3.40
	R	18	21 -84 21	3.59
Middle Temporal Gyrus	L	39	-45 -60 9	3.40*
Inferior Temporal Gyrus	L	21	-51 -12 -24	3.26
Parietal Lobe	R	5	27 -33 57	3.16

*significant at p <.05 after whole brain cluster correction with a t threshold of 2.86 and an extent of 53 voxels.

For completeness, peaks are reported for all clusters ≥ 15 voxels at p < .005 unc.

Table S5. Brain regions where higher BOLD responses were observed at the revelation of outcome of win trials in the instructed relative to the feedback session.

Region	***L/R***	***BA***	***MNI Coordinates***	***Z***
Temporal Lobe	R	37	45 -54 -12	3.98*
Medial Frontal Gyrus	L	32	-18 54 9	3.39
	R	32	18 51 9	3.63
DLPFC	L	46	-48 24 33	3.59*
Fusiform	R	37	36 -51 -21	3.49
Precentral Gyrus	R	3	48 -15 27	3.30
Frontal Lobe	R	9/46	30 15 21	3.26
Superior Frontal Gyrus	R	32	15 36 42	3.26
Parietal Lobe	L	40	-33 -36 45	3.23*

*Significant at p <.05 after whole brain cluster correction with a t threshold of 2.35 and an extent of 110 voxels.

For completeness, peaks are reported for all clusters ≥ 15 voxels at p < .005 unc.

Table S6. Areas where BOLD responses showed a stronger negative correlation with the activity in the seed area (DLPFC, Figure 5) during the win trials in the instructed session. Results were generated from a psychophysiological interaction (PPI) analysis (See materials and methods section for detail).

Region	*L/R*	*BA*	*MNI Coordinates*	*Z*
vmPFC	L	10	-6 48 -18	4.86*
Inferior Frontal Gyrus	R	38	27 24 -21	3.27
	R	38	39 15 -18	3.87*
Hippocampus	L	20	-24 -18 -24	3.77*
Middle Temporal Lobe	L	22	-51 -39 -3	3.63*
	L	21	-54 -18 -18	3.46*
Superior Frontal Gyrus	L	9	-6 60 30	3.28
Inferior Temporal Gyrus	R	21	45 0 -39	3.27
NAc	L	-	-3 6 -12	3.17*
Medial Frontal Gyrus	R	6	9 -18 66	3.19

*Significant at p <.05 after whole brain cluster correction with a t threshold of 2.07 and an extent of 182 voxels.

For completeness, peaks are reported for all clusters ≥ 15 voxels at p < .005 unc.

Supplementary Figures

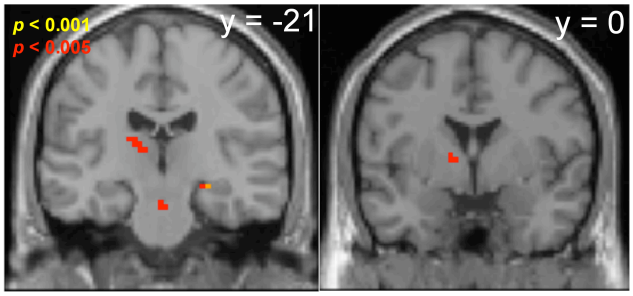

Figure S1. Direct contrast of prediction error (PE) related brain activities in feedback session and instructed session (feedback – instructed) showed greater involvement of BOLD activities in striatum, midbrain and hippocampus in the feedback session ($p < .005$ in red and $p < .001$ in yellow, unc).

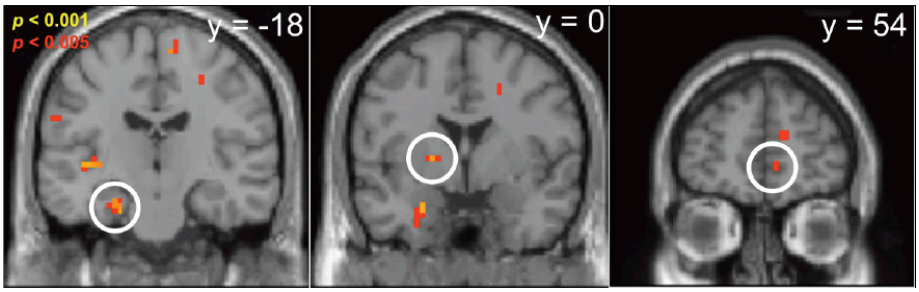

Figure S2. Brain regions that showed stronger negative connectivity to DLPFC (Figure 5*A*) in instructed than in feedback session ($p < .005$ in red and $p < .001$ in yellow, unc).

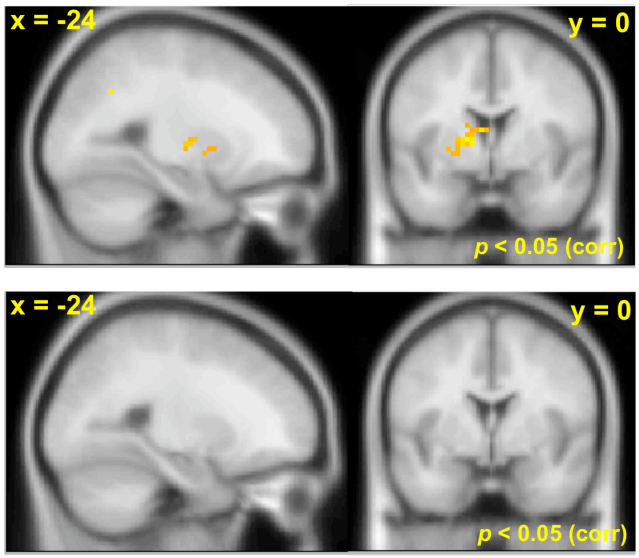

Figure S3. Brain regions that showed significant positive correlation with prediction errors derived from a simple RL model in the feedback (top) and instructed (bottom) sessions ($p < .05$ FDR corrected).

Reference:

1. Doll BB, Jacobs WJ, Sanfey AG & Frank MJ (2009) Instructional control of reinforcement learning: A behavioral and neurocomputational investigation. *Brain Res* 1299: 74-94.

2. Biele G, Rieskamp J & Gonzalez R (2009) Computational models for the combination of advice and individual learning. *Cog Sci* 33: 206-242.

3. Friston KJ, Worsley KJ, Frackowiak RSJ, Mazziotta JC & Evans AC (1993) Assessing the significance of focal activations using their spatial extent. *Hum Brain Mapp* 1: 210-220.

4. Hare TA, Camerer CF & Rangel A (2009) Self-control in decision-making involves modulation of the vmPFC valuation system. *Science* 324: 646-648.