Neuron, Volume *81*

Supplemental Information

Neural Computations Underlying Arbitration between Model-Based and Model-free Learning

Sang Wan Lee, Shinsuke Shimojo, and John P. O'Doherty

Supplemental Figures (Figure S1-S5)

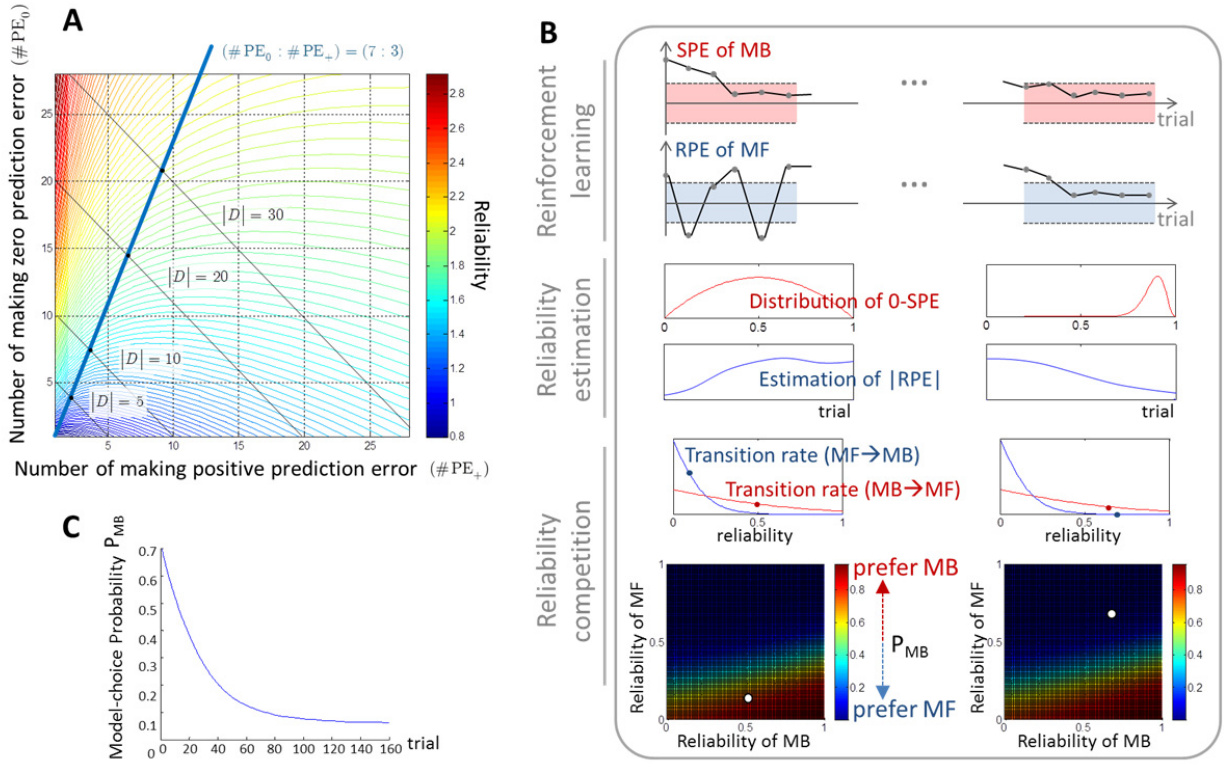

Figure S1, Related to **Figure 2** and **Experimental Procedures**. **(A)** Relationship between the reliability signals and prediction errors. Shown is the contour plot of reliability of a model's strategy as a function of the number of zero and positive prediction errors (PE). Only two kinds of PE (positive PE or zero PE) are considered for display. *D* is a set of PEs, $\#\mathrm{PE}_+$ refers to the number of positive PEs in *D*, $\#\mathrm{PE}_0$ refers to the number of zero PEs in *D*. $|D|$ is the cardinality of D ($|D| = \#\mathrm{PE}_+ + \mathrm{PE}_0$). Equi-reliability contours are color-coded. Equi-cardinality lines are in black. The thick blue line corresponds to the case where the ratio $(\#\mathrm{PE}_0 : \#\mathrm{PE}_+) = 7:3$. It illustrates the relationship between the PE and the reliability – ratio of PE and the number of observations. (1) Effect of the ratio between the amount of zero PE (#PE0) and the positive PE (#PE+). For example, if the model-based learning agent made a choice thirty times (|D|=30 line) and if twenty-one of them reported zero PE (#PE0=21) and nine of them reported positive PE (#PE+=9), then the corresponding reliability for the model-based learning agent would be around 2.2. However, if #PE0=9 and #PE+=21, then the reliability would be around 1.5. The more positive the PE, the lower the reliability becomes. (2) Effect of the number of observations. For example, if the model-based learning agent accumulated evidence with a constant ratio between $(\#\mathrm{PE}_0 : \#\mathrm{PE}_+) = 7:3$ (i.e., making a mistake with the fixed probability 0.3), then the reliability would increase with the number of

observations ($\left|D\right|$). The more observations, the more reliable your assessment becomes. **(B)** Example of our computational model's behavior during performance of the two-stage Markov decision task as described in (13), illustrating how the model choice probability (PMB) of our arbitration model is computed from PEs. It simply shows that the less PEs elicited by a specific learning strategy (the first row), the more reliable the model becomes (the second row), and the more weight for that model the arbitrator allocates (the third row). The red color code corresponds to the MB and the blue corresponds to the MF. The first row ("reinforcement learning") illustrates the change of the amount of state-prediction error (SPE) and reward prediction error (RPE) in the model-based learning system (MB) and the model-free (MF), respectively. The shaded area defines the range of zero PE (0-PE); the dotted lines define its upper and lower bound. The second row ("Bayesian inference") shows the corresponding posterior distribution of making 0-PE for each system. The graphs in the third row ("Two-state transition") show the transition rate as a function of the reliability, which is the mean to the variance of the posterior; each dot corresponds to the transition rate of the MB being pushed to the MF and vice versa. The shaded surface plot at the bottom shows the probability as a function of the reliability for the MB (x-axis) and for the MF (y-axis). Probability values are color-coded with the red being under the MB control and the blue being under the MF control. The white dots indicate the corresponding model choice probability (P_{MB}); the dot in the left plot means more control is assigned to the MB, whereas the right case means more control is assigned to the MF. **(C)** Model bias in the arbitration process aids transition from model-based to the model-free control. To validate the arbitration scheme we tested whether our computational model also works in a more restricted situation where the uncertainty of state-transitions is fixed. For this we ran our arbitration model, in which we deliberately removed the boundary condition that imposes a model choice bias, on the task used in (Gläscher et al., 2010). Shown is the probability of choosing the model-based system during the two-stage sequential Markov decision task used by (Gläscher et al., 2010). By optimizing the bias parameters ($A_\alpha, A_\beta, B_\alpha, B_\beta$; see **Supplemental Methods** - Dynamical transition model for reliability-based arbitration), we expected to find evidence of a gradual transition from MB to MF control as a function of training. We fit our model's free parameters to the behavioral data by minimizing the negative log-likelihood $-\sum \log P(s,a)$ of the choice *a* made, summed across all subjects and trials (the same procedure as the one of (Gläscher et al., 2010)). To have an overview of the model's behavior, we performed adaptive global optimization. We subsequently obtained a population of choice behaviors over 160 trials from the top ten best performing parameter fits. The plot shows the average timeline of the model choice probability (P_{MB}) of those models; x-axis refers to the trial and y-axis refers to the model choice probability. This averaged timeline exhibits the exponential transition from the MB to the MF as predicted, and it is consistent with behavioral accounts (Balleine and Dickinson, 1998). The same exponential decay effect was also exhibited for the model that was the overall best fit to the behavioral data.

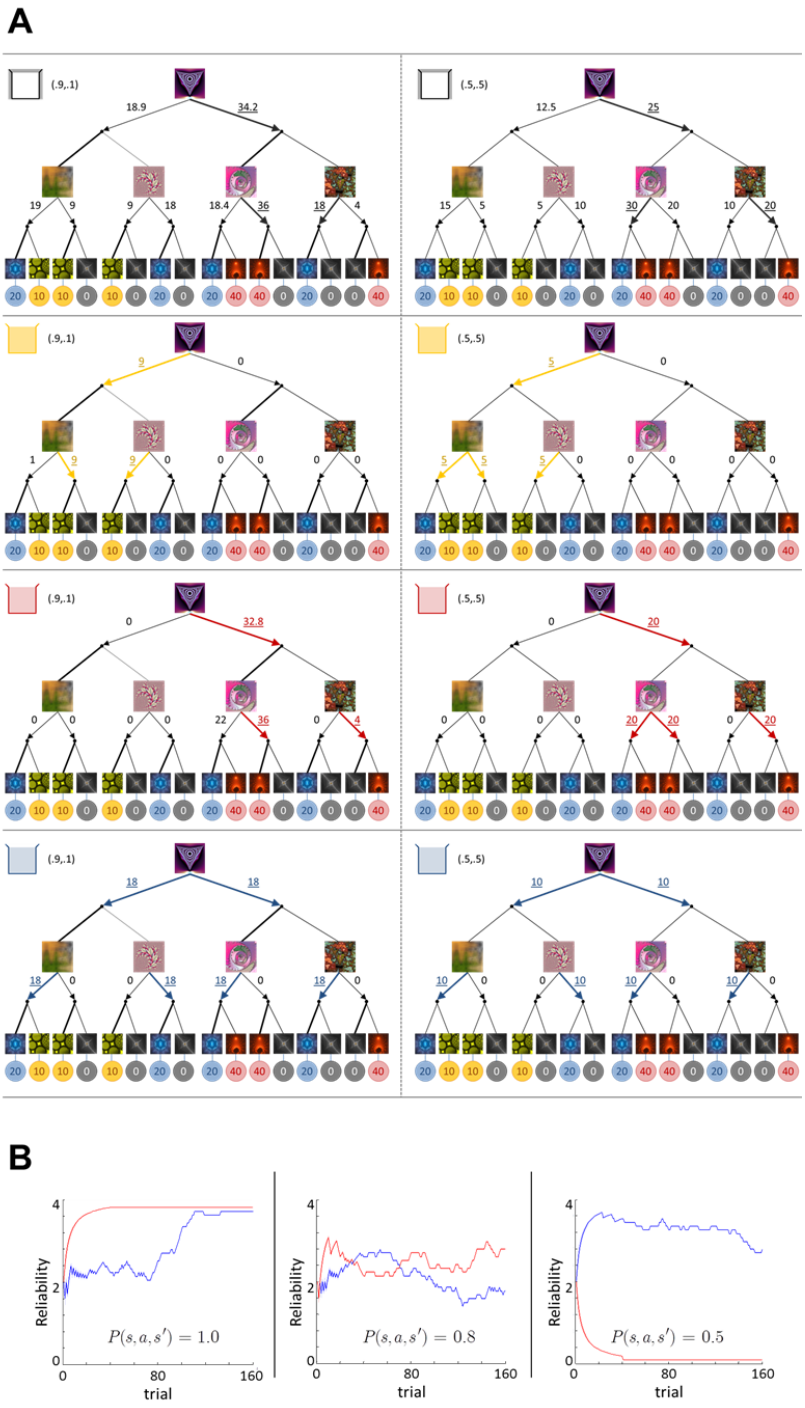

Figure S2, Related to **Figure 1** and **Experimental Procedures**. **(A)** Example of optimal choices during the Sequential two-choice Markov decision task, created by the backward induction of state-action values using a max strategy. Each fractal picture represents each state. Each arrow shows agent's choices, and the line thickness corresponds to the state transition probability. The color of the collecting box specifies

the goal assigned in each trial. The number next to each arrow represents the state-action value, which in this example is defined as the max of the next state values weighted by state-transition probability. **(B)** Effect of state-transition uncertainty on the performance profile of the model-based and the model-free system. To test whether our computational models perform as hypothesized, we examined a situation in which minimum, medium, and maximum uncertainties in state-transitions are considered. For a binary choice task, the minimum uncertainty condition corresponds to the state-transition probability (1,0), in which choices ensure transition to a certain state. On the other hand, the maximum uncertainty case corresponds to the state-transition probability (0.5,0.5), in which the chance of reaching a certain state given a particular choice is 0.5. Shown are the time courses of the reliability of the model-based system (red line) and the model-free system (blue line). The reliability is defined as the ratio of the posterior mean and the posterior variance for zero prediction error. In the minimum uncertainty condition where the state-transition probability $P(s,a,s') = 1.0$ given the current state *s*, binary choice *a*, and a particular state *s'*, the reliability of the model-based system is greater than the model-free system (shown in the leftmost), but it was the other way around in the maximum uncertainty condition ($P(s,a,s') = 0.5$, shown in the rightmost). Competitions are intense for the medium uncertainty case ($P(s,a,s') = 0.8$, shown in the middle). In a situation where state-transition uncertainties are minimal, the MB rule should ideally learn faster than the MF rule. Indeed, the reliability of the MB and the MF rules are fully consistent with this prediction (shown in the leftmost plot). In the maximum state-transition uncertainty situation, however, there would be a large amount of irreducible uncertainty left after learning is complete (Payzan-LeNestour and Bossaerts, 2011); the expectation is that the MB keeps generating a considerable amount of SPE, thus resulting in lower reliability for the MB system (shown in the rightmost plot). Taken together, this shows that we can deliberately make the reliability of one learning system greater than the other by changing the state-transition probability. The second and the third cases $P(s,a,s') = 0.8$ and $P(s,a,s') = 0.5$ correspond to our task design.

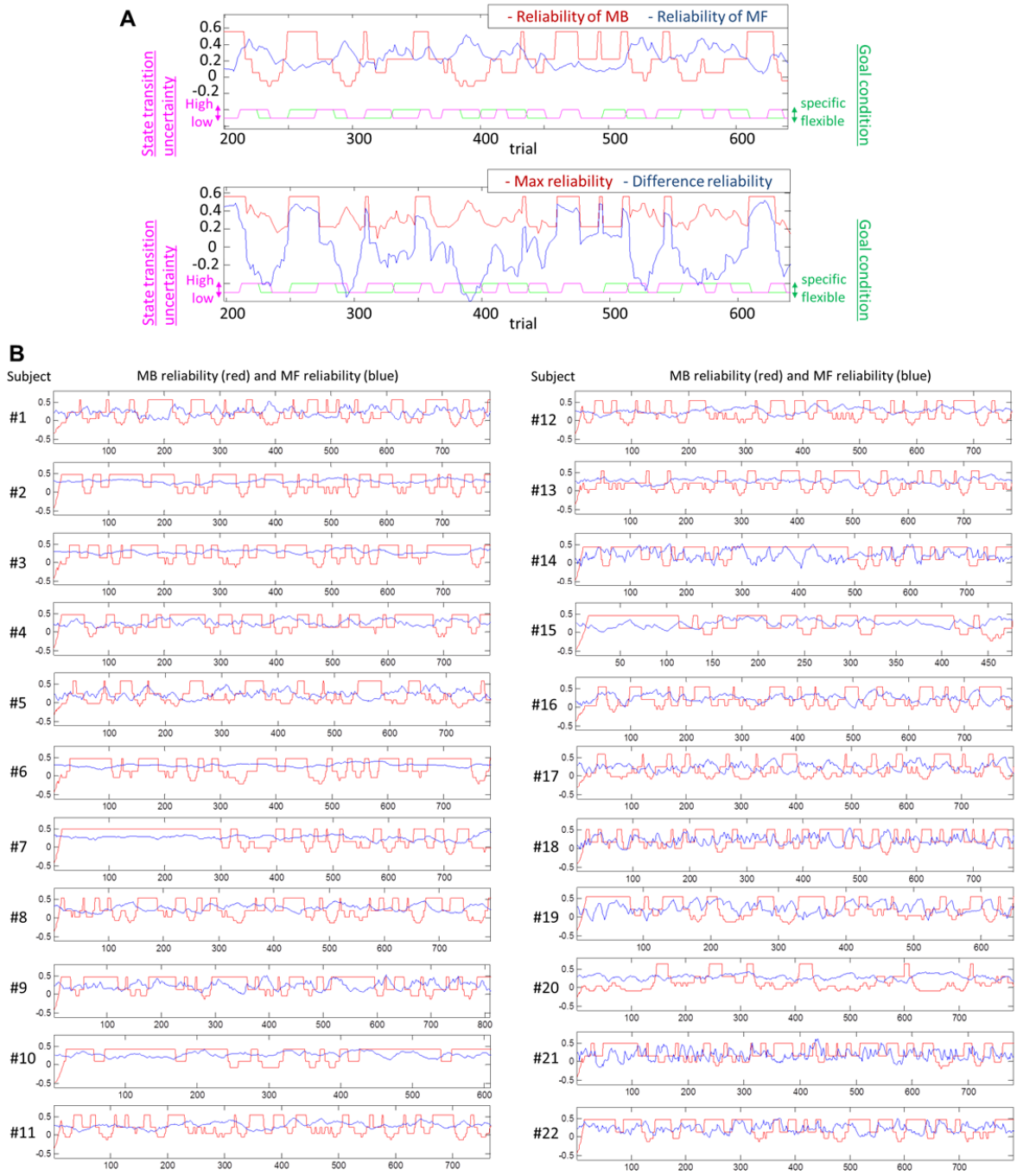

Figure S3, Related to **Figure 4**. Trial-by-trial reliability trace of the model-based and the model-free learning system. **(A)** Example of the trial-by-trial reliability trace of the model-based (red line) and the model-free (blue line) for a particular subject. The purple line shows the change of the state transition uncertainty, in which the up-state and the down-state corresponds to the high and the low uncertainty block, respectively. The green line shows the change of the goal condition, in which the up-state and the down-state corresponds to the specific and the flexible goal block, respectively. The reliability of the model-based tends to be high in the specific goal condition, whereas the reliability of the model-free tends

to be high in the flexible goal condition. The reliability of the model-based decreases with increases in the state-transition uncertainty, whereas the model-free is in favor during high state-transition uncertainty. **(B)** Trial-by-trial reliability trace for all trials and all subjects.

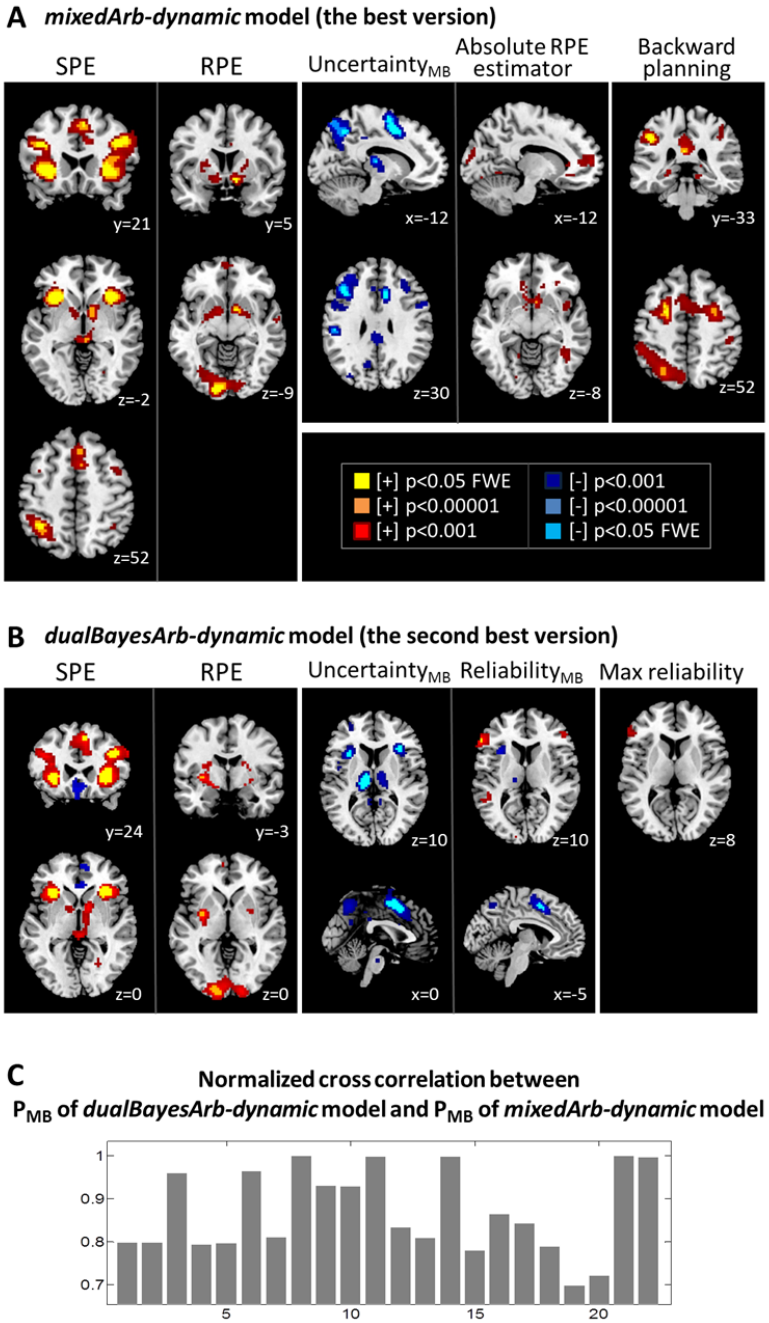

Figure S4, Related to **Figure 5**. **(A)** Imaging results from GLM analysis, for which the *mixedArb-dynamic* model is fit to each subject's choice data. (Left) State prediction error (SPE) responses are found in bilateral intraparietal sulcus, lateral prefrontal cortex, insula, globus pallidus extending to caudate. Reward prediction error (RPE) responses are found in the ventral and dorsal striatum. All activations survive familywise error correction (FWE p<0.05). (Middle) Significant effect for model-based uncertainty (Uncertainty$_{MB}$) in the dorsomedial prefrontal cortex, supplementary motor area, and thalamus (negative correlation, FWE p<0.05). Also found was a significant effect for the absolute reward prediction error

(Absolute RPE estimator) in the caudate (cluster-level corrected $p<0.05$). (Right) Activity in intraparietal sulcus, dorsolateral prefrontal cortex, insula, supplementary motor area, and posterior cingulate cortex is associated with the square-root of the updated amount of value signal in BACKWARD planning (FWE $p<0.05$), which quantifies the degree of computational demand put into backward planning in the model-based learning system. The yellow/orange/red and the cyan/light blue/blue color codes correspond to positive and negative correlations, respectively. Effects significant at $p<0.05$ (FWE corrected) are shown in yellow and cyan. The full list is shown in **Table S3**. **(B)** Imaging results from GLM analysis, for which the *dualBayesArb-dynamic* model is fit to each subject's choice data. Shown are the regions associated with the model's signals. Uncertainty and reliability of the model-free system ($Uncertainty_{MF}$, $Reliability_{MF}$) and the difference in reliability between the two systems did not survive corrected thresholds. The yellow/orange/red and the cyan/light blue/blue color codes correspond to positive and negative correlations, respectively. Effects significant at $p<0.05$ (FWE corrected) are shown in yellow and cyan. SPE and RPE signals were associated with activation in the same areas as in our main results with *mixedArb-dynamic*. The neural correlates of uncertainty and reliability of zero SPE mostly overlapped with the areas found in the *mixedArb-dynamic* analysis although the effects were weaker. Notably, we failed to identify locations associated with uncertainty of zero RPE (even at $p<0.001$ uncorrected), suggesting that the model-free system does not use a Bayesian mechanism for computing reliability because encoding uncertainty is an inherent feature of such a Bayesian computation. The fact that the model-free reliability and uncertainty signals are not present in this analysis motivated us to use the alternative arbitration model (*mixedArb-dynamic* model), in which a simpler mechanism for computing the degree of reliability of the RPE is used. **(C)** Correlation between the model choice probability(P_{MB}) of the two arbitrators – *dualBayesArb-dynamic* in which Bayesian estimation of reliability is applied to both the model-based and the model-free and *mixedArb-dynamic* in which Bayesian estimation of reliability is applied to the model-based and Pearce-Hall associability model is used to estimate the reliability of the model-free. *x*-axis refers to the subject index and *y*-axis indicates normalized cross correlation at zero lag. The *mixedArb-dynamic* exhibits different model choice pattern to the *dualBayesArb-dynamic* in eleven subjects (correlation value < 0.8)

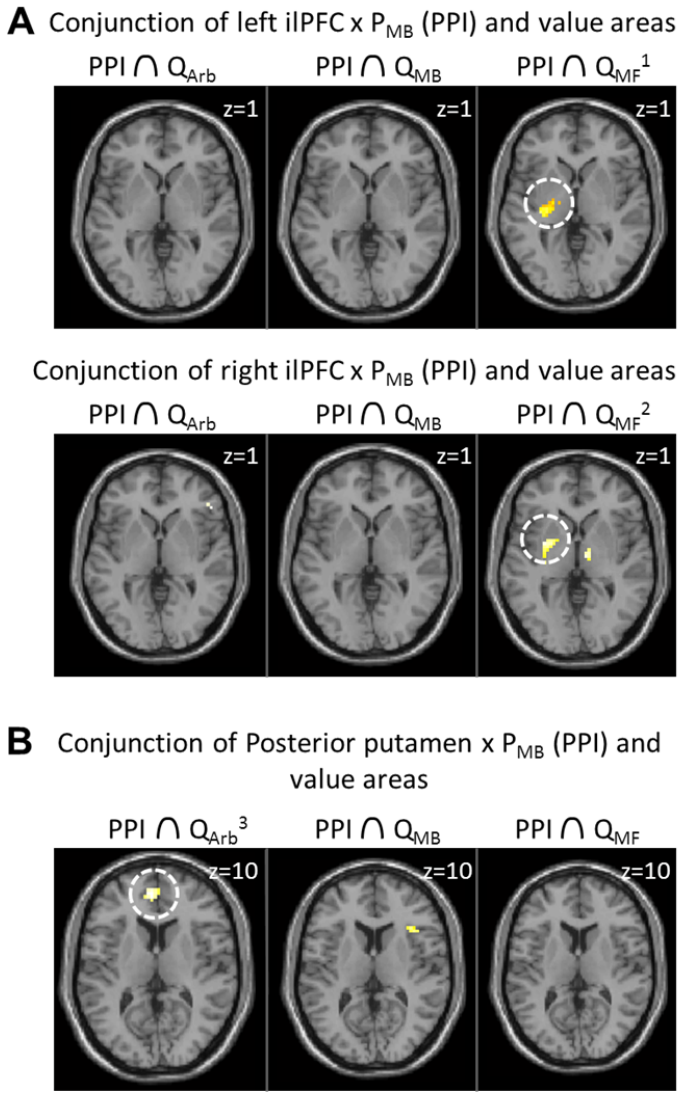

Figure S5, Related to **Figure 6** and **Figure 7**. Influence of the arbitration on the valuation systems. Shown are the conjunctions of the PPI result and the value areas. To conduct a comprehensive analysis, we applied an uncorrected p=0.001 threshold to individual contrasts and tested the value signals from the version shown in **Figure 6**. **(A)** Conjunction of the PPI result (the left and right inferior lateral prefrontal cortex x model choice probability PMB; **Figure 7A** and **Table S5**) and the value areas of the arbitrator, the model-based, and the model-free, respectively (QArb, QMB, QMF; **Figure 6**), showing that the PPI results (**Figure 7A**) can be ascribed primarily to coupling between the reliability area and the valuation area of the model-free system. 1, 2 : peaked at (-21,-13,1) and (-30,-1,7), respectively; survives small-volume correction within a 10-mm sphere centered on both coordinates (-33, -24, 0)(Tricomi et al., 2009) and (-27,-13,4)(Wunderlich et al., 2012). **(B)** Conjunction of the PPI result (the left posterior putamen x model choice probability PMB; **Figure 7B** and **Table S5**) and the value areas of the arbitrator, the model-

based, and the model-free, respectively (QArb, QMB, QMF; **Figure 6**), showing that the PPI results (**Figure 7B**) can be ascribed primarily to coupling between the reliability area and the valuation area of the arbitration system. 3 : vmPFC regions survive small-volume correction within a 10-mm sphere centered on all the coordinate (-3,42,-6) (Chib et al., 2009; Hare et al., 2011), and (-6,48,-8) (Boorman et al., 2009; Rushworth et al., 2011). All images are shown with threshold at $p<0.001$ for display purposes.

Supplemental Tables (Table. S1-S5)

Table S1. Performance comparison in terms of the trade-off between model fit and model complexity, Related to **Figure 2**.

Model	*MF alone*	*MB alone*	*dualBayesArb -mean*	*dualBayesArb -reliability*	*dualBayesArb -dynamic*	*mixedArb -mean*	*mixedArb -reliability*	*mixedArb -dynamic*
# param	2	2	4	4	6	4	4	6
AIC	1115.9	546.5	523.1	525.7	519.1	528.2	527.7	517.1*
AICc	1115.9	546.5	523.1	525.8	519.3	528.6	527.9	517.2*
BIC	1125.2	555.8	541.4	544.0	545.5	549.3	546.1	535.6*

The goodness of fit was compared between the models in terms of Akaike information criterion (AIC), AICc, and Bayesian information criterion (BIC), which each penalize for the number of free parameters in a slightly different manner. The test models include the model-based learning system (*MB alone*), the model-free learning system(*MF alone*), and the two types of arbitration models – *dualBayesArb* and *mixedArb*. *dualBayesArb-dynamic* refers to the full arbitration model in which Bayesian estimation of reliability is applied to both MB and MF, whereas *dualBayesArb-reliability* and *dualBayesArb-mean* refer to the reduced arbitration models w/o the biophysical two-state transition, in which the model choice probability is given by the ratio of the posterior mean and the variance (*dualBayesArb-reliability*) and by the posterior mean (*dualBayesArb-mean*), respectively. The equivalent of the original arbitrator proposed by Daw et al. (2005) is the *dualBayesArb-uncertainty*, the reduced arbitration model w/o biophysical two-state transition in which the model choice probability is given by the posterior uncertainty. *mixedArb-dynamic* refers to the full arbitration model in which Bayesian estimation of reliability is applied to MB and Pearce-Hall associability model is used to estimate the absolute value of MF's RPE, whereas *mixedArb-reliability* and *mixedArb-mean* refer to the reduced arbitration models w/o biophysical two-state transition in which the choice probability is given by the ratio of the posterior mean and the variance (*mixedArb-reliability*) and by the posterior mean (*mixedArb-mean*), respectively. The models are optimized for each individual. The median fitness value is shown due to the highly skewed performance distribution (non-Gaussian) and a few outliers. Asterisk(*) indicates the best performance for each criterion. In terms of BIC score, the best model is *mixedArb-dynamic*; it performs better than the second best model, *dualBayesArb-dynamic* (Wilcoxon signed rank test, $p<0.05$) and also better than the other models, including *MBalone* and the original arbitrator proposed by Daw et al., (2005) (Wilcoxon signed rank test, $p<0.01$). The performance of the Daw et al arbitration scheme is not shown here, although it performs less well than any of the *dualBayesArb* versions shown in the table.

Table S2. Estimated parameter values of *mixedArb-dynamic* model (the best version according to the performance comparison **Table S1**), Related to **Figure 3B** and **Figure 3C**.

parameter / subject	1	2	3	4	5	6
1	0.5684	0.1311	1.0268	5.5933	0.1407	0.1112
2	0.6984	0.0223	1.0000	1.3651	0.289	0.1477
3	0.7000	0.0246	1.0431	6.8312	0.3444	0.1498
4	0.6973	0.0716	1.0042	2.1552	0.3313	0.145
5	0.4779	0.1087	1.0056	1.3328	0.1694	0.0909
6	0.6860	0.0213	1.0033	6.9854	0.2808	0.0816
7	0.6580	0.0506	1.0009	5.6969	0.3255	0.0300
8	0.6003	0.0544	1.0055	9.7375	0.1385	0.1001
9	0.6975	0.1271	1.0000	3.704	0.3415	0.1036
10	0.6690	0.0493	1.0049	5.2908	0.3498	0.073
11	0.6047	0.0503	1.0002	9.5035	0.1157	0.1061
12	0.6950	0.0974	1.0148	1.0045	0.153	0.1991
13	0.5982	0.0568	1.1115	1.0661	0.1664	0.1485
14	0.6964	0.2095	1.0112	9.9334	0.0682	0.0301
15	0.6992	0.0898	1.0108	4.0382	0.2499	0.0516
16	0.5793	0.0987	2.8696	5.8722	0.1592	0.1015
17	0.5144	0.1129	1.0012	6.2778	0.2375	0.1283
18	0.6402	0.1878	1.0119	1.0392	0.3464	0.15
19	0.6105	0.1977	7.9023	8.1786	0.1887	0.0301
20	0.4037	0.0500	7.7070	2.2032	0.03	0.1858
21	0.6697	0.3281	1.0009	9.6769	0.1771	0.1303
22	0.6922	0.1752	1.0201	7.7754	0.0657	0.1252

Parameter: 1- the threshold for defining zero state prediction error, 2- learning rate for the estimate of absolute reward prediction error, 3- the amplitude of a transition rate function (MB→MF), 4- the amplitude of a transition rate function (MF→MB), 5- inverse softmax temperature, and 6- learning rate of the model-based and the model-free, respectively.

Table S3. Neural signatures of the model-based, the model-free, and the arbitration system signals, Related to **Figure 4** and **Figure S4A**.

x	y	z	Peak in region	Hemi	p (FWE)	Z-score
State prediction error (SPE : δ_{SPE})						
33	29	7	Insula	R	0.000	5.94*
-30	20	-2	Insula	L	0.000	5.87*
42	23	28	lPFC	R	0.000	5.71*
-39	8	25	lPFC	L	0.005	5.21*
-27	-73	31	IPS	L	0.004	5.28*
9	5	-2	Globus Pallidus	R	0.031	4.79*
Reward prediction error (RPE: δ_{RPE})						
9	5	-8	Ventral striatum	R	0.009	5.04*
-9	2	-8	Ventral striatum	L	0.004	4.02^{++}
-18	-4	-11	Amygdala	L	0.004	4.17^{++}
-24	5	10	Dorsal striatum (putamen)	L	0.004	3.51^{++}
27	-13	10	Dorsal striatum (putamen)	R	0.037	3.45^{+}
Variance of the posterior distribution of the model-based system making zero state prediction error (Uncertainty$_{MB}$: $Var\left(\theta_0\middle\vert_D\right)$) - negative correlation						
-9	5	49	dmPFC, SMA	L	0.000	6.61*
-51	-28	43	IPL	L	0.000	6.53*
42	-31	49	IPL	R	0.018	4.97*
-12	-19	10	Thalamus	L	0.050	4.71*
27	2	52	MFG	R	0.000	6.11*
-27	5	64	MFG	L	0.000	6.04*
-9	-58	46	Precuneus	L	0.007	5.20*
-39	23	31	dlPFC	L	0.009	5.15*
-30	20	4	Insula	L	0.023	4.91*
Estimate of absolute reward prediction error (absolute RPE estimator: Ω)						
6	11	-8	Caudate	R	0.029	4.55^{+}
-6	14	-8	Caudate	L	0.029	4.22^{+}
-9	59	7	vmPFC	L	0.039	4.07^{+}
Max reliability ($\max\left(\chi_{MB},\chi_{MF}\right)$)						
-54	38	3	ilPFC	L	0.001	5.52*
48	35	-2	ilPFC	R	0.001	5.48*
15	56	25	FPC	R	0.018	4.47^{+}

Difference reliability ($\chi_{MB} - \chi_{MF}$)						
3	32	10	ACC	R	0.019	4.13^{+}
Updated amount of value signal in BACKWARD planning ($\sqrt{\sum_{\forall s,a} \Delta Q(s,a)^2}$)						
-45	-40	46	IPS	L	0.000	6.18*
-24	5	55	dlPFC	L	0.017	4.93*
30	-1	52	dlPFC	R	0.010	4.31^{++}
48	-28	40	SMG	R	0.017	4.92*
-33	17	4	Insula	L	0.044	4.69*
-3	-34	25	PCC	L	0.050	4.65*
-39	26	25	IFC	L	0.011	4.41^{++}
-27	-31	-5	Hippocampus	L	0.050	4.41^{+}
18	-37	7	Hippocampus	R	0.004	4.19^{+}

- lPFC: lateral prefrontal cortex, ilPFC: inferior lateral prefrontal cortex, FPC: Frontopolar prefrontal cotex, IPS: Intraparietal sulcus, IPL: Inferior parietal lobule, MFG: Middle frontal gyrus, SMG: Supramarginal gyrus, IFC: Inferior frontal cortex, ACC: Anterior cingulate cortex, PCC: Posterior cingulate cortex.
- All the areas (marked with either "+", "++", or "*") survived after the whole-brain correction for multiple comparison at the cluster level (corresponding to "+"; height threshold t=3.53, extent>100 voxels), except for the value signals.
- p (FWE): corresponds to peak-level if the z-score indicates *, and corresponds to cluster-level if the z-score indicates ++/+.

*: thresholded p<0.05 FWE corrected at the peak-level, minimum 5 voxels extent.

$^{++}$: survives whole-brain correction for multiple comparison at the cluster level (height threshold t=3.53, extent>200 voxels).

$^{+}$: survives whole-brain correction for multiple comparison at the cluster level (height threshold t=3.53, extent>100 voxels).

Table S4. Neural representations of the value signals, Related to **Figure 6**.

x	y	z	Peak in region	Hemi	p (FWE)	Z-score
Chosen value of the model-based system only (Q_{MB})						
-3	38	-11	omPFC	L/R	0.033	3.16[1]
-3	41	-14	omPFC/ACC	L	0.042	2.99[2]
Chosen value of the model-free system only (Q_{MF})						
-9	8	55	SMA	L	0.001	5.58*
-48	26	34	dlPFC	L	0.003	5.37*
46	23	43	dlPFC	R	0.007	4.35[++]
9	35	40	dmPFC	R	0.000	4.57[++]
-36	56	22	alPFC	L	0.013	4.16[++]
-27	-4	1	Posterior putamen	L	0.010	3.52[3]
Regions with the average effect of the model-based and the model-free's chosen value ($Q_{MB\|MF}$)						
-12	-1	67	SMA	L	0.038	4.76*
12	32	37	dmPFC	R	0.001	4.44[++]
Value difference of the arbitration system – positive correlation (chosen-unchosen)						
-9	29	-11	vmPFC/OFC	L	0.050	4.69*
Value difference of the arbitration system – negative correlation (unchosen-chosen)						
33	20	7	Insula	R	0.037	4.78*
-36	23	-2	Insula	L	0.033	4.11[+]
-9	17	46	SMA	L	0.000	4.68[++]

- omPFC: Orbital & medial prefrontal cortex, dl/dm/alPFC: Dorsolateral/Dorsomedial/Anteriorlaternal prefrontal cortex.

- p (FWE): peak-level if the corresponding z-score indicates *, cluster-level if the z-score indicates ++/+.

*: threshold p<0.05 FWE corrected at the peak-level, minimum 5 voxels extent.

[++]: survives whole-brain correction for multiple comparison at the cluster level (height threshold t=3.53, extent>200 voxels).

[+]: survives whole-brain correction for multiple comparison at the cluster level (height threshold t=3.53, extent>100 voxels).

[1]: survives small-volume correction within a 10-mm sphere centered on the coordinate both (0,32,-13) and (3,32,-17) (Wunderlich et al., 2012).

[2]: survives small-volume correction within a 10-mm sphere centered on the coordinate both (-6,44,-5) (Hare et al., 2011).

[3]: survives small-volume correction within a 10-mm sphere centered on the coordinates (-27,-13,4) and 20-mm on (-33, -24, 0) ((Wunderlich et al., 2012) and (Tricomi et al., 2009), respectively).

Table S5, Modulatory interactions of the arbitration system (PPI analysis), Related to **Figure 7.**

x	y	z	Peak in region	Hemi	p (FWE)	Z-score
Negative modulation of left inferior lateral prefrontal cortex mediated by model choice probability(P_{MB})						
63	14	-2	STG	R	0.025	4.85*
-45	5	22	IFG	L	0.029	4.81*
-48	29	4	IFG	L	0.031	4.79*
3	-4	40	MCC	R	0.000	3.98^{+}
-27	-19	4	Posterior putamen	L	0.001	4.38^{1}
Negative modulation of right inferior lateral prefrontal cortex mediated by model choice probability(P_{MB})						
3	-10	7	Thalamus	R	0.042	4.76*
-27	8	4	Putamen	L	0.000	4.48^{++}
-3	32	16	ACC	L	0.021	4.12^{+}
-36	-22	-8	Posterior putamen	L	0.004	3.84^{1}
Negative modulation of right frontopolar cortex mediated by model choice probability(P_{MB})						
51	11	-11	STG	R	0.017	4.20^{+}
33	-10	1	Posterior putamen	R	0.003	4.00^{2}
Negative modulation of left posterior putamen mediated by model choice probability(P_{MB})						
0	62	10	vmPFC	L/R	0.021	4.88*
-24	56	28	FPC	L	0.000	4.24^{++}
Negative modulation of supplementary motor area mediated by model choice probability(P_{MB})						
-21	59	31	FPC	R	0.005	4.38^{+}
Negative modulation of dorsal medial prefrontal cortex mediated by model choice probability(P_{MB})						
48	50	-5	OFC	R	0.005	4.38^{+}
24	59	1	OFC	R	0.005	4.23^{+}

- STG: Superior Temporal Gyrus, IFG: Inferior Frontal Gyrus, MCC: Middle cingulate cortex, ACC: Anterior cingulate cortex, STG: Superior Temporal Gyrus, vmPFC: Ventromedial prefrontal cortex, FPC: Frontopolar prefrontal cotex.

- Left inferior lateral prefrontal cortex (seed region): BOLD signal was extracted from 5-mm sphere centered on (-54,38,3). Psychological factor: P_{MB} (parametric). To show the regions functionally overlapping with Q_{MF}, we added an inclusive mask by regions associated with Q_{MF}.

- Right inferior lateral prefrontal cortex (seed region): BOLD signal was extracted from 5-mm sphere centered on (48,35,-2). Psychological factor: P_{MB} (parametric). To show the regions functionally overlapping with Q_{MF}, we added an inclusive mask by regions associated with Q_{MF}.

- Right frontopolar cortex (seed region): BOLD signal was extracted from 5-mm sphere centered on (15,56,25). Psychological factor: P_{MB} (parametric). To show the regions functionally overlapping with Q_{MF}, we added an inclusive mask by regions associated with Q_{MF}.
- Left posterior putamen (seed region): BOLD signal was extracted from 5-mm sphere centered on (-27,-4,1). Psychological factor: P_{MB} (parametric).
- Supplementary motor area (seed region): BOLD signal was extracted from 5-mm sphere centered on (-9,8,55). Psychological factor: P_{MB} (parametric).
- Dorsal medial prefrontal cortex (seed region): BOLD signal was extracted from 5-mm sphere centered on (9,35,40). Psychological factor: P_{MB} (parametric).

- p (FWE): peak-level if the corresponding z-score indicates *, cluster-level if the z-score indicates ++/+.

*: thresholded p<0.05 FWE corrected at the peak-level, minimum 5 voxels extent.

$^{++}$: survives whole-brain correction for multiple comparison at the cluster level (height threshold t=3.53, extent>200 voxels).

$^{+}$: survives whole-brain correction for multiple comparison at the cluster level (height threshold t=3.53, extent>100 voxels).

^{1}: survives small-volume correction within a 10-mm sphere centered on both coordinates (-33, -24, 0)(Tricomi et al., 2009) and (-27,-13,4)(Wunderlich et al., 2012).

^{2}: survives small-volume correction within a 10-mm sphere centered on both coordinates (33, -24, 0) and (-27,-13,4) (flipped from (Tricomi et al., 2009) and (Wunderlich et al., 2012), respectively).

Supplemental Methods

Model-free (MF) and Model-based (MB) reinforcement learning. The model-free SARSA learner (MF) (Sutton and Barto, 1998) incorporates a temporal-difference learning rule, by which the reward prediction error (RPE) δ_{RPE} is computed and the corresponding state-action value $Q_{\mathrm{MF}}(s,a)$ for the action *a* in the state *s* is updated:

$$\delta_{RPE} = r(s') + \gamma Q_{\mathrm{MF}}(s',a') - Q_{\mathrm{MF}}(s,a),$$
$$\triangle Q_{\mathrm{MF}}(s,a) = \alpha \delta_{RPE},$$

where s, s' refers to the current and the next state, respectively, a, a' refers to the action in the current state and in the next state, respectively, $r(s')$ denotes the obtained reward in state s', γ is a temporal discount factor fixed at 1 because the two-step task does not allow subjects to choose between rewards at different delays (Gläscher et al., 2010), α denotes the free parameter that controls the model's learning rate.

We implemented the model-based learner (MB), which is equipped with FORWARD learning (following our previous study (Gläscher et al., 2010)) and BACKWARD planning. The FORWARD learning component uses experience with state transitions to update a state-transition matrix $T(s,a,s')$ of transition probabilities, which represents the probability of the agent arriving at the state s' if it made a choice a in state s. Whenever the agent transitions from one state to another, the state prediction error (SPE) is computed and the corresponding state-action value is updated:

$$\delta_{SPE} = 1 - T(s,a,s'),$$
$$\triangle T(s,a,s') = \eta \delta_{SPE},$$
$$Q_{\mathrm{MB}}(s,a) = \sum\nolimits_{s'} T(s,a,s') \left\{ r(s') + \max\nolimits_{a'} Q_{\mathrm{MB}}(s',a') \right\},$$

where η denotes the free parameter associated with learning rate.

The update rule is based on a dynamic programming scheme (from the Bellman optimality equation (Sutton and Barto, 1998)). Note that the first term of the SPE is set to 1 to incorporate the assumption that the state space is deterministic. As such, SPE is always positive. The limiting value of SPE quantifies the degree of irreducible uncertainty left even after learning is complete (Payzan-LeNestour and Bossaerts, 2011).

In addition, the model-based learning system is also capable of doing BACKWARD planning whenever an agent is presented with an explicit goal (e.g., change in a specific goal condition or transition from the flexible to the specific goal condition):

$$r(s)\begin{cases} = R & for\ a\ goal\ state, \\ = 0 & otherwise. \end{cases}$$

$for\ i = 3, 2,$

  $for\ s \in S_{i-1}$

    $Q_{\mathrm{MB}}(s,a) = \sum_{s'} T(s,a,s')\left\{ r(s_i) + \max_{a'} Q_{\mathrm{MB}}(s',a') \right\},\ for\ all\ a.$

  end

end

where *R* is the reward value corresponding to the goal state, S_i refers to the set of states in *i*-th stage. This backward planning corresponds to the case where, for example, “Now that I have a yellow collecting box, the only way I receive the reward is to find a yellow coin ($r(s_{\mathrm{Yellow}}) = 40$); the red and the blue coins are not valuable anymore ($r(s_{\mathrm{Red}}) = r(s_{\mathrm{Blue}}) = 0$). I can see how to get to the room that ensures the highest chance of receiving a yellow coin (using the backward update rule) because I know how the rooms are connected ($T(s,a,s')$).” This step is akin to repeating the FORWARD update process for all possible states and actions.

Bayesian Reliability estimation of MB and MF strategy. We use a simple hierarchical empirical Bayes approach to compute the reliability of a learning strategy given the history of the prediction error (PE). PE refers to SPE for the case of MB and RPE for the case of MF. Specifically, we (1) used the conditional probability of the model-based system making zero state-prediction error, (2) assumed that the parameter is drawn from a Dirichlet prior, and (3) used conjugacy when performing Bayesian inference.

First, the conditional probability distribution of making positive, zero, and negative PE is given by

$$P(\mathrm{PE} \mid \theta) = \begin{cases} \theta_1 & if\ \mathrm{PE} < -\omega\ (\mathrm{PE}_1) \\ \theta_2 & if\ \mathrm{PE} > +\omega\ (\mathrm{PE}_2) \\ \theta_0 & otherwise \qquad (\mathrm{PE}_0) \end{cases},$$

where w refers to the tolerance level and $\sum_{i=0}^{2} \theta_i = 1$. Here $\theta_0, \theta_1, \theta_2$ represents the probability of making zero, negative, and positive prediction error, respectively. $\mathrm{PE}_0, \mathrm{PE}_1, \mathrm{PE}_2$ refers to the case in which a certain event leads to zero, negative, and positive prediction error, respectively. The two free parameters (w; one for SPE and the other for RPE) determines the extent to which subjects tolerate prediction error. There is no negative threshold for SPE because SPE is positive by definition.

The tolerance threshold defines the graininess of the determination that the prediction error is positive, negative or zero.

In order to keep track of our belief about θ , we assume the following prior distribution $P(\theta)$:

$$\left(\theta_0,\theta_1,\theta_2\right) \sim \mathit{Dirichlet}(\lambda_1,\lambda_2,\lambda_3),$$

where $\sum_{i=0}^{2}\lambda_i = 1$.

Suppose that we experienced *T* discrete events that causes a set of corresponding prediction errors

$$D = \{\mathrm{PE}(1),\mathrm{PE}(2),\cdots,\mathrm{PE}(T)\}.$$

Then the posterior $\theta\big|_D$ is also *Dirichlet* distribution (conjugacy):

$$\mathit{Dirichlet}(\lambda_0 + \#\mathrm{PE}_0,\lambda_1 + \#\mathrm{PE}_1,\lambda_2 + \#\mathrm{PE}_2),$$

where $\#\mathrm{PE}_i\,(i = 0,1,2)$ refers to the number of occurrence of the event that leads to PE_i.

The expectation and the variance of the posterior is given as follows:

$$E\left(\theta_j\big|_D\right) = \left(1 + \#\mathrm{PE_j}\right)\Big/\left(3 + \left|D\right|\right),\ j = 0,1,2 \text{ and}$$

$$Var\left(\theta_j\big|_D\right) = \frac{\left(1 + \#\mathrm{PE}_j\right)\left(2 + \sum_{j\neq i}\#\mathrm{PE}_i\right)}{\left(3 + \left|D\right|\right)^2\left(4 + \left|D\right|\right)},\ j = 0,1,2,$$

where $\left|D\right| = \sum_{i=0}^{2}\#\mathrm{PE}_i$, the cardinality of D.

This posterior distribution shows the degree of goodness or badness of the current learning strategy. For example, the expectation of $E\left(\theta_0\big|_D\right)$ quantifies the degree of the current strategy making zero prediction error, and the variance $Var\left(\theta_0\big|_D\right)$ quantifies the belief about making zero prediction error. The consistency of the events one experiences determines skewedness and peakedness of the posterior distribution (Koller and Friedman, 2009); the skewedness can be represented as an expectation, and the peakedness can be represented as a variance.

We now quantify the reliability of the learning strategy:

$$\chi = \chi_0\Big/\sum_{i=0}^{2}\chi_i\,.$$

where $\chi_i = E\left(\theta_i\big|_D\right)\Big/Var\left(\theta_i\big|_D\right),\ i = 0,1,2.$

Note that this ratio is the inverse of the index of dispersion (or called Fano factor for windowed data), known to measure the reliability with which the performance can be estimated from a time window that

collects events (Pennini and Plastino, 2010) and known to characterize uncertainty in a communication channel (Janesick, 2000); it has also been used to quantify the efficiency of information transfer in population of neurons (Ma et al., 2006). The reliability is a function of not only the number of non-zero PEs but also the total number of observations. The defining characteristics of the reliability are that it increases with the number of corresponding prediction errors while penalizing inconsistencies in those observations. For example, it increases with the number of observations given the fixed ratio of events that cause non-zero PE to zero PE (**Figure S1A**).

Pearce-Hall associability for reliability estimation. An alternative method of non-Bayesian estimation of MF reliability is to use a Pearce-hall type associability rule to substitute for the Bayesian update (Li et al., 2011; Le Pelle, 2004; Sutton, 1992), based on the unsigned reward prediction error (Krugel et al., 2009).

The update of the absolute RPE estimator Ω is given by

$$\Delta\Omega = \eta\big(\left|\mathrm{RPE}\right| - \Omega\big),$$

where η denotes the constant free parameter that controls the model's learning rate. Here we simply define the reliability as $\chi_{\mathrm{MF}} = \big(\mathrm{RPE}_{\max} - \Omega\big)\big/\mathrm{RPE}_{\max}$ with $\mathrm{RPE}_{\max}$ being the upper bound of RPE ($\mathrm{RPE}_{\max} = 40$). The update of the reliability is thus

$$\Delta\chi_{\mathrm{MF}} = \eta\Big\{\Big(1 - \left|\mathrm{RPE}\right|\big/\mathrm{RPE}_{\max}\Big) - \chi_{\mathrm{MF}}\Big\}.$$

The reliability becomes zero if the agent predicts the maximum amount of RPE ($\Omega \to \left|\mathrm{RPE}\right| = \mathrm{RPE}_{\max}$), whereas it reaches the maximum if the agent predicts zero RPE ($\Omega \to 0$).

Let us consider the case where an agent performs a two-stage Markov decision task with the state-transition probability fixed at (0.7,0.3) (akin to (Gläscher et al., 2010)). In the early stages of the task, MF would keep producing a large amount of RPEs while MB would quickly adapt to the environment that elicits a decreasing SPE. The corresponding MF reliability quickly decreases, whereas the MB reliability gradually increases. This would result in a situation in which the MF becomes less reliable than the MB (shown in the second row, left of **Figure S1B**). However, in the late stages of the task, both the MF and the MB strategies would converge to the optimal state-action value. Since both models generate a small amount of PEs consistently, the corresponding reliability would quickly increase. This would cause almost equal reliability for both learning systems (shown in the second row, right of **Figure S1B**).

Dynamical transition model for reliability-based arbitration. Next we implemented a push-pull mechanism to govern how the reliability-based competition between MB and MF mediates value computation. For this we introduced a dynamical two-state transition model inspired by biophysical

neuronal models (Dayan and Abbott, 2001). We consider a state associated with the probability p_{MB} that the control is allocated to MB strategy and the other state associated with the probability $p_{\mathrm{MF}} = 1 - p_{\mathrm{MB}}$ for choosing the MF strategy.

The transition rate α (a transition MF→MB) is a function of the summary statistics of the posterior probability of the Bayesian model:

$$\alpha\left(\chi_{\mathrm{MF}}\right) = \frac{A_{\alpha}}{1 + \exp\left(B_{\alpha}\chi_{\mathrm{MF}}\right)},$$

where A_{α}, B_{α} represents the maximum transition rate and the steepness, respectively. The transition function copes with the situation that the control tends to be passed to MB strategy (i.e., α increases) if the MF reliability is weak.

Likewise, the transition rate β (a transition MB→MF) is given as follows:

$$\beta\left(\chi_{\mathrm{MB}}\right) = \frac{A_{\beta}}{1 + \exp\left(B_{\beta}\chi_{\mathrm{MB}}\right)}.$$

A boundary condition is imposed to incorporate a bias so that all being equal control will pass from the MB to the MF over time; this is meant to accommodate the fact that the habits tend to emerge with increased training (Balleine and Dickinson, 1998; Gläscher et al., 2010). They are

$$B_{\alpha} = \log\left(\alpha(1)^{-1}A_{\alpha} - 1\right),\ B_{\beta} = \log\left(\beta(1)^{-1}A_{\beta} - 1\right),$$

with $A_{\alpha} \geq 2\alpha(1),\ A_{\beta} \geq 2\beta(1)$. These conditions are used as a constraint when optimizing the model. In all simulations, the boundary condition is fixed at $\alpha(1) = 0.1,\ \ \beta(1) = 0.01,$ which were generated as a result of an out of sample model fit to an independent dataset (Gläscher et al., 2010).

The change of the state value is given by the difference between the inward current and the outward current:

$$\frac{dp_{\mathrm{MB}}}{dt} = \alpha(1 - p_{\mathrm{MB}}) - \beta p_{\mathrm{MB}} \ \Leftrightarrow\ \tau_n \frac{dp_{\mathrm{MB}}}{dt} = p_{\mathrm{MB},\infty} - p_{\mathrm{MB}},$$

where $\tau_n = \frac{1}{\left(\alpha + \beta\right)},\ p_{\mathrm{MB},\infty} = \frac{\alpha}{\left(\alpha + \beta\right)}$. The p_{MB} is set to 0.8 whenever there is a transition from the flexible goal condition to the specific goal condition and 0.2 from the specific to the flexible goal condition. The parameters were optimized using an independent dataset in a preliminary analysis (using data from (Gläscher et al., 2010)). However, we also found that this is not a critical parameter; the results do not

change with any initial value in the range from 0.7 to 0.9 for a transition from the flexible goal condition to the specific goal condition and with any value from 0.1 to 0.3 for the other case.

Finally, the state-action value is given by the weighted average of the state-action value of MB and MF system.

$$Q(s,a) = p_{\mathrm{MB}} Q_{\mathrm{MB}}(s,a) + (1 - p_{\mathrm{MB}}) Q_{\mathrm{MF}}(s,a).$$

For example, full transfer to the model-free strategy (P_{MB}=1) indicates that the value signal of the MB system solely determines the final state-action value.

The alternative hypothesis "the weights are always one or zero" was ruled out in our preliminary study. To test this, we optimized the arbitration model with the following value integration:

$$Q(s,a) = \left(\left(p_{\mathrm{MB}} Q_{\mathrm{MB}}(s,a) \right)^{v} + \left(p_{\mathrm{MF}} Q_{\mathrm{MF}}(s,a) \right)^{v} \right)^{1/v},$$

where v denotes the degree of integration. $v = \infty$ (or >10 given the upper bound of Q-value) means the value integration occurs in a "winner-take-all" fashion (in other words, the weight is a binary variable). The optimized parameter is around 4, rejecting the alternative hypothesis $v \geq 10$ (t-test p<1e-5).

Given the state-action value, the arbitration model finally selects actions stochastically according to the following softmax function (Gläscher et al., 2010; Luce, 1959):

$$P(s,a) = \frac{\exp(\tau Q(s,a))}{\sum_b \exp(\tau Q(s,b))},$$

where τ is the "inverse temperature parameter controlling the extent to which the agent made a choice with the higher valued action.

Another characteristic feature of the model is reciprocity - one switches to a new strategy either because the current strategy has not been working or because an alternative strategy sounds more tempting. Note that with a zero time constant and an equalized bias this model converges to a trivial case, in which the probability of choosing the MB is purely determined by the Bayesian model. Let us follow the scenario of the previous section where the MF becomes less reliable than the MB in the early stage of the task. In the two-state transition model, the corresponding transition rate of MF→MB would be higher than of MB→MF, resulting in having a high probability of choosing the MB ($P_{\mathrm{MB}} > 0.5$) (shown in the third row, right of **Figure S1B**). On the other hand, if both the MF and the MB became equally reliable, the corresponding transition rate of MF→MB would be smaller than of MB→MF due to the model bias imposed on this transition model. This lowers the probability of choosing the MB ($P_{\mathrm{MB}} < 0.5$) (shown in the third row, right of **Figure S1B**), and thus model-free/habitual control is more dominant, as found in mammalian behavior (Balleine and Dickinson, 1998; Daw et al., 2005).

Parameter Estimation. The free parameters in these models are (*1*) a threshold for defining zero state prediction errors, *(2)* a learning rate for the estimate of absolute reward prediction error, *(3)* the amplitude of a transition rate function (MB→MF), *(4)* the amplitude of a transition rate function (MF→MB), *(5)* inverse softmax temperature, and *(6)* a learning rate of the model-based/model-free; we use a single learning rate parameter for the model-based and the model-free ($\eta = \alpha$) because the two models' performance difference are stable for a given learning rate that guarantees convergence.

The first two parameters, *(1)* and *(2),* provide a clear intuition about how the assessment of the reliability of the model-based and the model-free is sensitive to the trial-by-trial prediction error. The first parameter, the threshold for defining the zero state prediction error, and the second parameter, the learning rate for the estimate of absolute reward prediction error, represents how subjects are vulnerable to a mistake about predicting the reward and the state, respectively.

The next two parameters, *(3)* and *(4)*, provides an intuition about the bias that are used to determine the innate preference of the model. They determine the extent to which each subject prefer the model-based or the model-free learning strategy.

The remaining parameters, *(5)* and *(6)*, are well-known: the decision parameter (inverse softmax temperature) and the learning rate of the model-based and the model-free, all of which have been dealt with many literatures and thus the necessity is unquestionable.

We used the Nelder-Mead simplex algorithm (Lagarias et al., 1998) to estimate the parameters by minimizing negative log-likelihood $-\sum \log P(s,a)$ of the obtained choices *a* given the observed choices and rewards, summed over all trials for each subject. To minimize the risk of finding a local but not global optimal solution, we ran optimization 100 times with randomly generated seed parameters. The optimized parameters are shown in **Table S2**.

Model-Comparison. In order to compare the performance of different possible arbitration schemes on explaining participant's behavior on the task we used Akaike information criterion(AIC)(Akaike, 1974), AICc(Burnham and Anderson, 2002), and Bayesian information criterion(BIC)(Schwarz, 1978) to correct for the number of free parameters used in each model.

For the model comparison, we tested the following variants of the arbitration:

A. *dualBayesArb*: the arbitration model in which Bayesian estimation of reliability is applied to both MB and MF (see **Supplemental Methods** - Bayesian Reliability estimation of MB and MF strategy for more details)

(i) *dualBayesArb-dynamic*: arbitration with a dynamical transition process, which incorporates model bias so that all else being equal model-free control is favored (see **Supplemental Methods** - Dynamical transition model for reliability-based arbitration for more details).

(ii) *dualBayesArb-reliability*: an arbitration process without the dynamical transition, in which reliability directly and instantaneously controls the model choice probability for both MB and MF.

(iii) *dualBayesArb-mean*: arbitration without dynamical transition, in which the mean of the posterior directly and instantaneously controls the model choice probability for both MB and MF.

B. *mixedArb*: arbitration model in which the Bayesian estimation of reliability is applied to MB and a Pearce-Hall asssociability-like rule is used to estimate the absolute value of MF's RPE. (see **Supplemental Methods** - Bayesian Reliability estimation of MB and MF strategy and Pearce-Hall associability for reliability estimation for more details)

(i) *mixedArb-dynamic*: the arbitration with dynamical transition, which incorporates model bias so that all else being equal model-free control is favored (see **Supplemental Methods** - Dynamical transition model for reliability-based arbitration for more details).

(ii) *mixedArb-reliability*: arbitration without dynamical transition, in which reliability directly and instantaneously controls the model choice probability for MB; reliability assessment for MF is the same as (i).

(iii) *mixedArb-mean*: the arbitration without biophysical transition, in which the mean of the posterior directly and instantaneously controls the choice probability for MB; reliability assessment for MF is the same as (i).

C. *MB alone*: the model-based learning system alone (see **Supplemental Methods** - Model-free (MF) and Model-based (MB) reinforcement learning for more details)

D. *MF alone*: the model-free learning system alone (see **Supplemental Methods** - Model-free (MF) and Model-based (MB) reinforcement learning for more details)

Note that the model (i) becomes (ii) when the time constant of the dynamical transition is zero (corresponding to an instantaneous model), the model (ii) becomes (iii) when the mean substitutes for the reliability, and all the models in (A) and (B) reduce to (C) or (D) when it is endowed with a zero time constant of the dynamical transition and a fixed model bias towards MB and MF, respectively.

The above six different types of arbitration strategies (A-(i),(ii),(iii) and B-(i),(ii),(iii)) allows us to systematically test a variety of alternative arbitration schemes. They were used to answer the following questions as to how different types of reliability contribute to the arbitration process. The results of the model comparison are shown in **Table S1**.

Question1: Is reliability competition a dynamic process?

-- hypothesis 1-a. the reliability competition is an *instantaneous* process, in which the behavioral adaptation (PMB) is based on a short-term estimation of reliability. The corresponding models are (1) *dualBayesArb–mean*, (2) *dualBayesArb–reliability*, (3) *mixedArb–mean*, and (4) *mixedArb–reliability*.

-- hypothesis 1-b. the reliability competition is a *dynamic* process, in which the behavioral adaptation is modulated by the intensity of competition between the MB and the MF systems. They are (1) *dualBayesArb–dynamic* and (2) *mixedArb–dynamic*.

The model-fits support hypothesis 1b, given that the dynamic form for the arbitrator performs better in accounting for participant's choices than the instantaneous form.

Question2: Is the reliability computation implemented via the average amount of prediction error, uncertainty of prediction error, or both?

-- hypothesis 2-a. the reliability computation is based on the average amount of prediction errors. They are (1) *dualBayesArb–mean* and (2) *mixedArb–mean*.

-- hypothesis 2-b. the reliability computation is based on the uncertainty in the amount of prediction errors. The performance of such models is significantly worse than any of the *dualBayesArb* models listed in **Table S1**. This includes Daw et al.'s model(Daw et al., 2005), in which the model choice probability is given by the posterior variance of the state-action value.

-- hypothesis 2-c. the reliability computation is based on the ratio of the posterior mean and the variance. They are (1) *dualBayesArb–reliability* and (2) *mixedArb–reliability*.

The model-fits provide support for hypothesis 2c as models using the ratio of the posterior mean reliability and variance in realibility accounted better for participant's behavior than did any of the other models.

Question3. Does the arbitrator use different methods to assess the reliability of model-based and model-free learning?

-- hypothesis 3-a. the brain computes the reliability of the model-free system in the same way as the model-based. They are (1) *dualBayesArb–mean*, (2) *dualBayesArb–reliability*, and (3) *dualBayesArb–dynamic*.

-- hypothesis 3-b. the brain computes the reliability of the model-free system using a reliability estimate based on an approximation of reliability generated by taking the absolute value of the reward-prediction error akin to the Pearse-Hall learning rule. . These model variants are (1) *mixedArb–mean*, (2) *mixedArb–reliability*, and (3) *mixedArb–dynamic*.

The model-fits provide support for hypothesis 3b, in that the model versions using the absolute RPE approximation for the model-free reliability signal outperforms the full Bayesian version.

Taking all of these model-fitting procedures together therefore, the best fitting model was *mixedArb-dynamic*, which is the model we use primarily in the Results. This model fit significantly better than the next best-fitting model (Wilcoxon signed rank test at $p<0.05$; **Table S1**). This model uses a ratio of mean predicted prediction error to variance of the predicted prediction error as the model-based reliability signal, uses a dynamical arbitration process and incorporates an absolute-valued RPE signal as the means of

computing the reliability within the model-free system. The arbitrator models described above also outperformed the version of arbitration proposed by Daw et al.(Daw et al., 2005). Furthermore, as can be seen from **Table S1**, the arbitration model accounted better for subjects' behavior than either the model-based (C) or the model-free alone (D).

Interaction between model parameters. To show that there is little chance that the free parameters of the arbitration model trade off against or interact with each other, we measured correlation for all parameter combinations on parameter sets from the top 15 arbitration models. Specifically, we computed the coefficient of determination for every combination of the parameter sets for each individual subject. First, we ran the global optimization process to find the optimal parameter sets, and then selected the top 15 parameter fits for each individual subject; each of the parameter sets corresponds to a local optimum.

Significant correlation would mean that there is an interaction between parameters, by which the different combination of parameter sets would exhibit similar model fits or the change in one parameter value is compensated for the loss of model fit due to the change in another parameter value. We found that the coefficient value is very small across subjects (mean 0.16, standard deviation 0.17), suggesting that there is little interaction between the parameters.

Implementation of *mixedArb-dynamic* and *dualBayesArb-dynamic* at the neural level. In order to provide a full description of how *mixedArb-dynamic* model and *dualBayesArb-dynamic* model is implemented at the neural level, respectively, we tested the following signals:

A. PE signal of MB/MF (SPE/RPE): Shown in **Figure S4A** (*mixedArb-dynamic*) and **Figure S4B** (*dualBayesArb-dynamic*), respectively.

B. Performance assessment of MB (Uncertainty of 0-SPE): Shown in **Figure S4A** (*mixedArb-dynamic*) and **Figure S4B** (*dualBayesArb-dynamic*), respectively.

C. Performance assessment of MF: We failed to identify locations associated with uncertainty of zero RPE (Uncertainty of 0-RPE) for *dualBayesArb-dynamic* model (even at p<1e-3 uncorrected). However, we found locations associated with |RPE| for *mixedArb-dynamic* model (**Figure S4A**).

D. Reliability of MB/MF

(i) Reliability of MB: Shown in **Figure 4A** (*mixedArb-dynamic*) and **Figure S4B** (*dualBayesArb-dynamic*), respectively.

(ii) Reliability of MF: We failed to identify locations associated with the reliability of MF for *dualBayesArb-dynamic* model (even at p<1e-3 uncorrected). However, we found locations for *mixedArb-dynamic* (**Figure 4A**).

(iii) Max Reliability: Shown in **Figure 4A** (*mixedArb-dynamic*) and **Figure S4B** (*dualBayesArb-dynamic*), respectively.

(iv) Difference Reliability: We failed to identify locations associated with the difference reliability for *dualBayesArb-dynamic* model (even at p<1e-3 uncorrected). However, we found locations for *mixedArb-dynamic* (**Figure 4A**).

E. Planning of MB: Shown in **Figure S4A** (*mixedArb-dynamic*). The results for *dualBayesArb-dynamic* are the same as for the backward planning of the *mixedArb-dynamic*.

GLM design. The general linear model (GLM) was used to generate voxelwise statistical parametric maps (SPMs) from the fMRI data. We created subject-specific design matrices containing the following regressors: (R1) regressors encoding the average BOLD response at two choice states and one outcome states, (R2,R3) two parametric regressors encoding the model-derived prediction error signals - state prediction error (SPE) of MB and reward prediction error (RPE) of MF (refer to Section - Model-free (MF) and Model-based (MB) reinforcement learning), (R4) a parametric regressor encoding the uncertainty of zero SPE given a set of observations *D*:

$$Var\left(\theta_0\big|_D\right).$$

We included this in the GLM matrix because this is one of the components that enters into the computation of reliability, which is the ratio of the mean over the variance of the prediction error estimate and is also tested in the GLM matrix. (R5) a parametric regressor encoding the estimate of absolute RPE (refer to Section - Pearce-Hall associability for reliability estimation), (R6) a parametric regressor encoding max or difference reliability of MB and MF (refer to Section - Bayesian Reliability estimation of MB and MF strategy):

$$\max\left(\chi_{MB},\chi_{MF}\right),\ \chi_{MB}-\chi_{MF},$$

(R7,R8) two parametric regressors encoding the chosen value of the model-based and the model-free system, respectively (Q_{MB} and Q_{MF}), (R9) one parametric regressor encoding the chosen minus the unchosen value of the arbitration system (Q_{Arb}). We tested both the chosen values and chosen minus unchosen signals for the output of the arbitration system, as such models have been reported in prior literature in this region suggesting both types of signals are often present (see e.g. (Boorman et al., 2009; Rushworth et al., 2011)). The chosen minus the unchosen signal showed stronger correlation; why this is the case is still unclear, but one possibility is that it relates to effects of attention (see (Lim et al., 2011) for a detailed elaboration of this hypothesis). (R10) one parametric regressor encoding the updated amount of all the state-action values of the model-based system after the BACKWARD planning (refer to in Section 1.2.1):

$$\sqrt{\sum_{\forall s,a}\left\{Q_{new}(s,a)-Q_{old}(s,a)\right\}^2},$$

and (R11) a nuisance partition containing six regressors that encoded the movement displacement as estimated from the affine part of the image realignment procedure. Each regressor explains the nested

process in our computational hypothesis; (R2,R3) correspond to low-level learning systems, (R4,R5,R6) correspond to the components of the arbitrator that acts on those prediction errors, (R7,R8) are the value signals dictated by the learning systems, (R9) are the value comparator of the arbitration controller, and finally, (R10) corresponds to the component that correlates with value update. Definitions of the parametric regressors (R2)-(R10) are given in the previous sections. Note that the GLM analysis was conducted without serial orthogonalization of the parametric regressors in order to avoid effects of orthogonalization order complicating interpretation of the results.

Note that all our main neural findings (SPE, RPE, uncertainty of MB, absolute RPE estimator, max reliability, difference reliability signal, chosen value of MB, chosen value of MF, and chosen-unchosen value signal of the arbitration system) are all based on the same GLM. We use a different GLM for the PPI analysis.

Statistical threshold for whole brain analysis. A single basic statistical threshold (whole-brain cluster correction) was used throughout with the exception of SVC correction for some a-priori predicted brain areas. However, many of the activations survived a more stringent whole-brain FWE corrected threshold at the single-voxel level at $p<0.05$. This includes SPE, RPE and most importantly, our reliability signals. The areas surviving this stringent threshold are listed in **Table S3** and **Table S4** with '*' marked and also shown in **Figure 4** and **Figure S4** (cyan and yellow blobs in the statistical maps). Areas surviving after the whole-brain correction for multiple comparisons at the cluster level are listed in **Tables S3** and **Tables S4** with a '+' marked. '++' denoted areas with effects stronger than '+'; specifically, '+' refers to the case where the region survived after cluster-level correction with the height threshold $t=3.53$ (extent>=100 voxels), whereas '++' was used with the height threshold $t=3.53$ (extent>200 voxels). For the areas (posterior putamen and vmPFC) about which we had a strong a priori hypothesis about value signals (Boorman et al., 2009; Chib et al., 2009; Hare et al., 2011; Rushworth et al., 2011; Tricomi et al., 2009; Wunderlich et al., 2012), small volume corrections were performed within a 10-mm sphere; areas surviving this less-stringent yet still statistically appropriate procedure are listed in **Table S4** and **Table S5**. In addition, for all the figures, in order to show the full extent of the activations we used a unitary stratification: $p<0.05$ FWE, $p<1e-5$ uncorrected, and $p<1e-3$ uncorrected.

Supplemental References

Akaike, H. (1974). A new look at the statistical model identification. IEEE Trans. Automat. Contr. *19*, 716–723.

Balleine, B.W., and Dickinson, A. (1998). Goal-directed instrumental action: contingency and incentive learning and their cortical substrates. Neuropharmacology *37*, 407–419.

Boorman, E.D., Behrens, T.E., Woolrich, M.W., and Rushworth, M.F.S. (2009). How Green Is the Grass on the Other Side? Frontopolar Cortex and the Evidence in Favor of Alternative Courses of Action. Neuron *62*, 733–743.

Burnham, K.P., and Anderson, D.R. (2002). Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach (Springer-Verlag).

Chib, V.S., Rangel, A., Shimojo, S., and O'Doherty, J.P. (2009). Evidence for a common representation of decision values for dissimilar goods in human ventromedial prefrontal cortex. J. Neurosci. *29*, 12315–12320.

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.

Dayan, P., and Abbott, L.F. (2001). Theoretical Neuroscience – Computational and Mathematical Modeling of Neural Systems (MIT press).

Gläscher, J., Daw, N.D., Dayan, P., and O'Doherty, J.P. (2010). States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. Neuron *66*, 585–595.

Hare, T. a, Schultz, W., Camerer, C.F., O'Doherty, J.P., and Rangel, A. (2011). Transformation of stimulus value signals into motor commands during simple choice. Proc. Natl. Acad. Sci. *108*, 18120–18125.

Janesick, J.R. (2000). Scientific Charge-Coupled Devices (International Society for Optical Engineering).

Koller, D., and Friedman, N. (2009). Probabilistic Graphical Models – Principles and Techniques (MIT press).

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

Lagarias, J.C., Reeds, J.A., Wright, M.H., and Wright, P.E. (1998). Convergence Properties of the Nelder-Mead Simplex Method in Low Dimensions. SiAM J. Optim. *9*, 112–147.

Li, J., Schiller, D., Schoenbaum, G., Phelps, E.A., and Daw, N.D. (2011). Differential roles of human striatum and amygdala in associative learning. Nat. Neurosci. *14*, 1250–1252.

Lim, S.L., O'Doherty, J.P., and Rangel, A. (2011). The Decision Value Computations in the vmPFC and Striatum Use a Relative Value Code That is Guided by Visual Attention. J. Neurosci. *31*, 13214–13223.

Luce, R.D. (1959). Individual choice behavior: a theoretical analysis (New York: Wiley).

Ma, W.J., Beck, J.M., Latham, P.E., and Pouget, A. (2006). Bayesian inference with probabilistic population codes. Nat. Neurosci. *9*, 1432–1438.

Payzan-LeNestour, E., and Bossaerts, P. (2011). Risk, Unexpected Uncertainty, and Estimation Uncertainty: Bayesian Learning in Unstable Settings. PLoS Comput. Biol. *7*.

Le Pelle, M.E. (2004). The role of associative history in models of associative learning: a selective review and a hybrid model. Q. J. Exp. Psychol. B *57*.

Pennini, F., and Plastino, A. (2010). Diverging Fano factors. J. Phys. Conf. Ser. *246*.

Rushworth, M.F.S., Noonan, M.P., Boorman, E.D., Walton, M.E., and Behrens, T.E. (2011). Frontal Cortex and Reward-Guided Learning and Decision-Making. Neuron *70*, 1054–1069.

Schwarz, G.E. (1978). Estimating the dimension of a model. Ann. Stat. *6*, 461–464.

Sutton, R.S. (1992). Adapting Bias by Gradient Descent: An Incremental Version of Delta-Bar-Delta. In Proceeding of Tenth National Conference on Artificial Intelligence, (MIT press),.

Sutton, R.S., and Barto, A.G. (1998). Reinforcement Learning (MIT press).

Tricomi, E., Balleine, B.W., and O'Doherty, J.P. (2009). A specific role for posterior dorsolateral striatum in human habit learning. Eur. J. Neurosci. *29*, 2225–2232.

Wunderlich, K., Dayan, P., and Dolan, R.J. (2012). Mapping value based planning and extensively trained choices in the human brain. Nat. Neurosci. *15*, 786–791.