Neuron, volume *74*

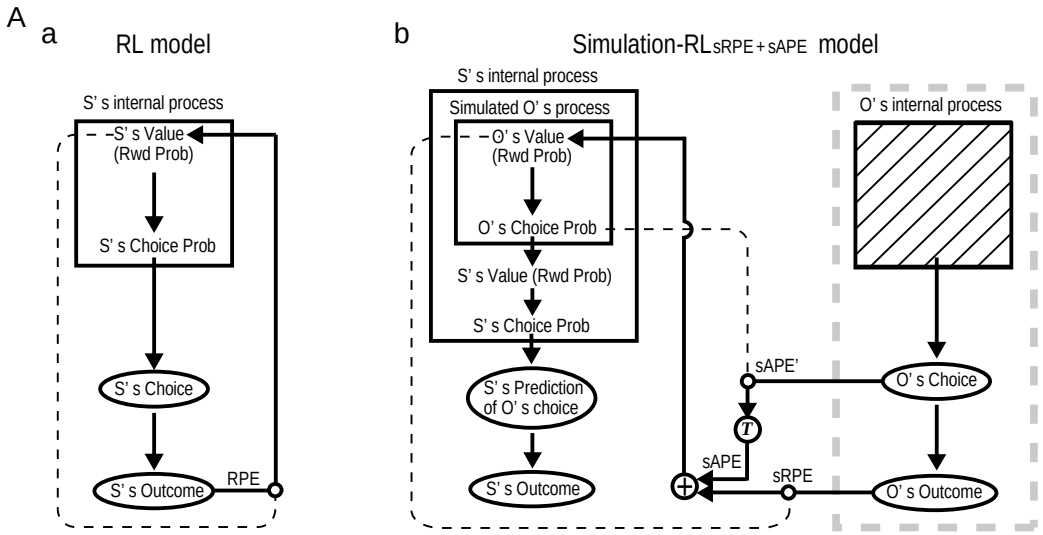
Supplemental Information

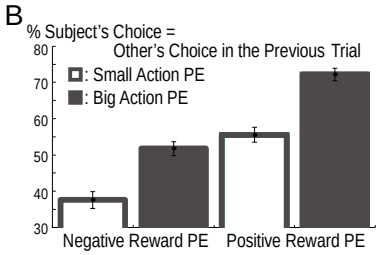
Learning to Simulate Others' Decisions

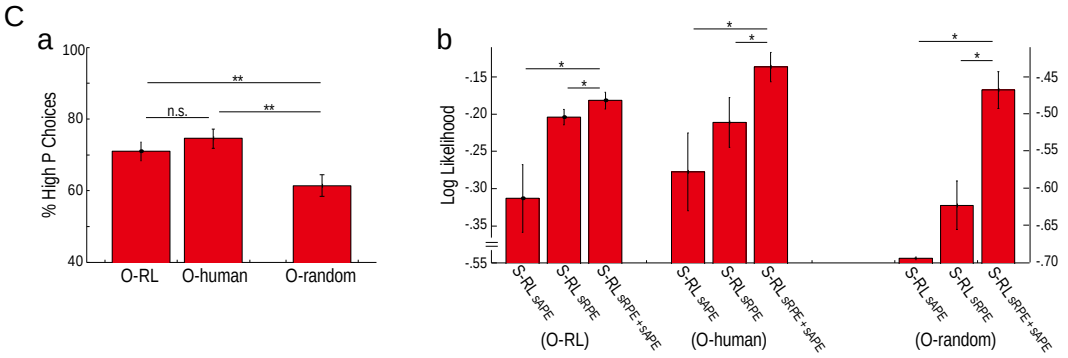
Shinsuke Suzuki, Norihiro Harasawa, Kenichi Ueno, Justin L. Gardner, Noritaka Ichinohe, Masahiko Haruno, Kang Cheng, and Hiroyuki Nakahara

Inventory of Supplemental Information

Supplemental Figures







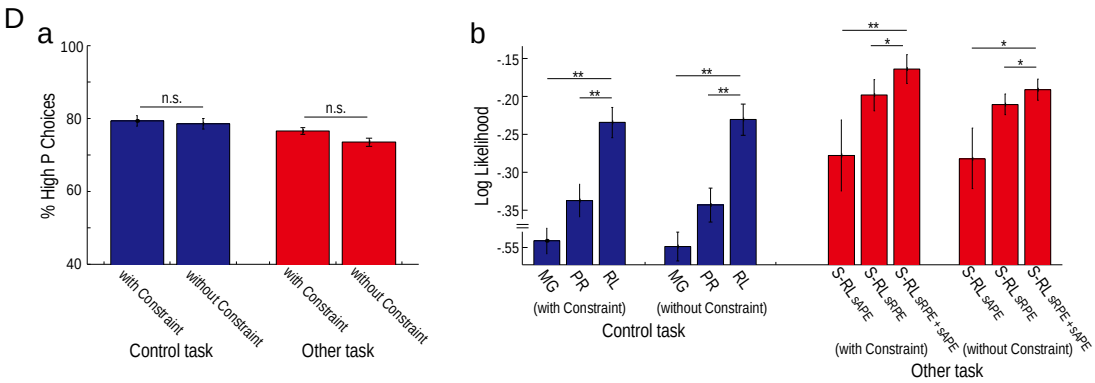


Figure S1 – related to Figure 1: Schematic diagrams of decision making processes used in this study and additional behavioral results

(A) Schematic diagram for value-based decision making processes in both the Control and Other tasks based on a reinforcement learning (RL) model (a, b, respectively). (a) Box indicates the subjects' (S's) internal decision making process. As modeled by the RL, at the time of decision, subjects use the learned values of options to generate the choice probability of the stimulus, and accordingly make a choice decision. When the outcome is presented, the value of the chosen option (or the stimulus reward probability) is updated, using reward prediction error (RPE: discrepancy between S's value and actual outcome). (b) Decision making process of subjects during the Other task is modeled by Simulation-$RL_{sRPE+sAPE}$ (S-$RL_{sRPE+sAPE}$) model. The large box on the left indicates the subject's internal process; the smaller box inside indicates the other's (O's) internal decision making process being simulated by the subject. The large box on the right, outlined by a thick dashed line, corresponds to what the other is 'facing in this task,' and is equivalent to what subjects were facing in the Control task (compare with the schematic in (a)). The hatched box inside corresponds to the other's internal process, which is hidden from the subjects. As modeled by the S-$RL_{sRPE+sAPE}$, at the time of decision, subjects use the learned simulated-other's value to first generate the simulated-other's choice probability (O's Choice Prob), based on which they generate their own value (S's Value) and the subject's choice probability for predicting the other's choice (S's Choice Prob). Accordingly, subjects then predict the other's choice. Once the outcome is shown, subjects update the simulated-other's value using the simulated-other's reward and action prediction errors (sRPE and sAPE), respectively; sRPE is the discrepancy between the simulated-other's value and the other's actual outcome, and sAPE is the discrepancy between the simulated-other's choice probability and the other's actual choice, in the value level. The simulated-other's action prediction error is first

generated in the action level (denoted by sAPE' in the figure) and transformed (indicated by *T* in the open circle) to the value level, becoming the sAPE to update the simulated-other's value, together with the sRPE.

(B) Effects of simulated-other's reward and action prediction errors on subjects' choice behavior on the next trials during the fMRI experiment. We show the mean percentages (±SEM) of times (across subjects; n=36) that the subject's prediction of the other's chosen option in the next trial coincided with the other's chosen option in the previous trial in each of the four cases: when the reward prediction error is negative (two left bars) or positive (two right bars), and when the action prediction error is smaller (open bars) or larger (filled bars) than the median.

(C) Subjects' behavior when the other's choices were generated by risk-neural RL (O-RL), risk-neutral humans (O-human), or a random-chooser (O-random). The results of this additional experiment support the rationale for the use of the fitted risk-neutral RL model in the main report. (a) Mean percentages (±SEM) of choosing the stimulus with the higher reward probability (across subjects; n=17); shown as the averages of all trials. Asterisks above the horizontal lines indicate significant differences between the indicated means (**P<0.01; two-tailed paired t-test; n.s., non-significant as P> 0.05). The subjects behaved similarly, regardless of whether the other's choices were generated by the O-RL or an O-human, but they behaved differently when the other's choices were randomly generated. Although not shown in the panel, here we note a baseline result; the O-RL-generated *other's choices* of the stimulus with the higher reward probability were not significantly different from the O-human-generated other's choices (P > 0.05, two-tailed paired t-test), but were significantly different from the O-random-generated other's choices (P < 0.001). (b) Models' fit to behaviors. Each bar (±SEM) indicates the log likelihood of each model, averaged over subjects and normalized by the number of trials (thus a larger magnitude indicates a better fit to behavior). *P<0.05, one-tailed

paired *t*-test over AIC distributions. The comparison indicates that S-$RL_{sRPE+sAPE}$ model best fit all three choice conditions (O-RL, O-human and O-random). Abbreviations for each model are the same in Figure 1D. See Supplemental Experimental Procedures for further details of this experiment.

(D) Subjects' behavior with and without an additional constraint on reward magnitude randomization. The results of this additional experiment demonstrate that the subjects' behaviors in both the Control and Other tasks did not significantly differ with or without the additional constraint. (a) Mean percentages (±SEM) of choosing the stimulus with the higher reward probability (across subjects; n=21) with and without the constraint are shown in the same format as panel C; n.s., non-significant as $P > 0.05$. Blue for the Control task and red for the Other task. In both tasks, the subjects' behaviors were not significantly different under the two conditions. (b) Models' fit to behaviors in the Control (*left*) and Other (*right*) tasks. We show the log likelihood of each model in the same format as panel C (b); $*P<0.05$ $**P<0.01$, one-tailed paired *t*-test over AIC distributions. In both tasks, the best fitted model reported in the main text was also the best fitted model in the condition without the constraint. See Supplemental Experimental Procedures for further details of this experiment.

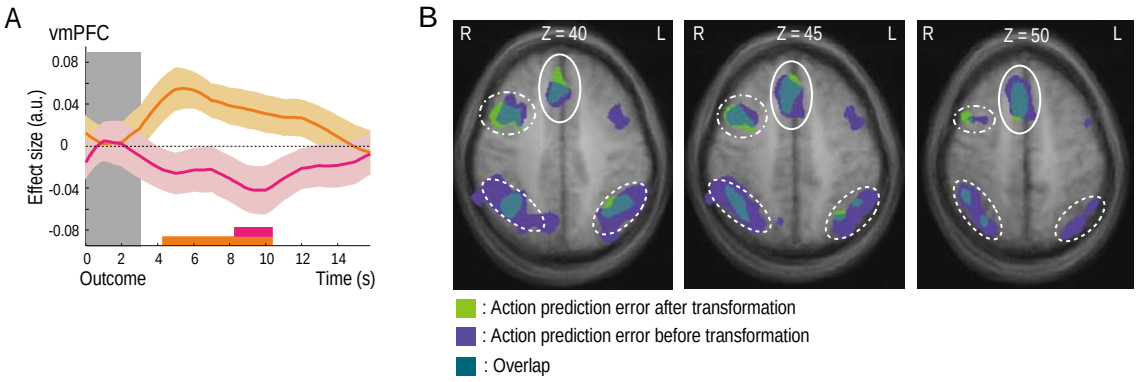


Figure S2 related to Figure 2: Additional results of the neural correlates of simulated other's reward and action prediction errors

(A) Time course of the component parts of the neural correlates of the simulated-other's reward prediction error in the vmPFC. Time course of effect sizes of the other's reward outcome (orange) and the simulated-other's reward probability (pink); the corresponding colored shading indicates ±SEM (n=36). To generate this plot, we first defined an ROI in the vmPFC based on the BOLD signals that were significantly correlated with the simulated-other's reward prediction error (Figure 2A). To investigate the two components of the error (the simulated-other's reward prediction error equals "the other's reward outcome (1 if the other-chosen stimulus is the rewarded stimulus, or 0 otherwise)" minus "the simulated-other's reward probability"), we transformed the BOLD signals in the ROI into z-scores over trials for each subject. Each time slice had a 200-ms resolution starting at, and aligned to, the onset of the OUTCOME phase and ended 16 s later. We then performed a first-order linear regression, "z-scored BOLD signals = a Other's Reward Identity + b Simulated-Other's Reward Probability" in each time slice for each subject. The mean and SEM of the estimated coefficients (effect sizes) for a and b over subjects were then plotted (orange and pink curves correspond to a and b, respectively). Gray shading indicates the OUTCOME duration. The

thick horizontal lines in the corresponding colors at the bottom indicate the periods during which the effect was significantly different from zero ($p < 0.05$, *t*-test). The transformation to z-scores mentioned above was employed so that the effect sizes could be compared among different time slices. We used a one-tailed *t*-test to examine the significance of the effect size in each time slice against the null hypothesis that it equaled zero.

(B) Neural activity significantly modulated ($P < 0.05$, corrected) by the action prediction error in two levels. Activity modulated by the 'after-transformed' (in value level) action prediction error (green), by the untransformed (in the action level) action prediction error (purple), and by the overlap of two activations (dark blue); the action prediction error in the action level significantly modulated BOLD signals ($P < 0.05$, corrected) in the dorsomedial prefrontal cortex (dmPFC; TAL x=6, y=26, z=46), the right dorsolateral prefrontal cortex (dlPFC; x=30, y=8, z=46), and the bilateral temporoparietal junction and posterior superior temporal sulcus (TPJ/pSTS; x=45, y= -52, z=43 and x=-39, y=-67, z=46), in addition to some other significantly modulated areas. The solid, dotted-dashed, and dashed ovals highlight the overlap in the dmPFC, the right dlPFC and the TPJ/pSTS, respectively. The maps are thresholded at $P < 0.005$, uncorrected for display.

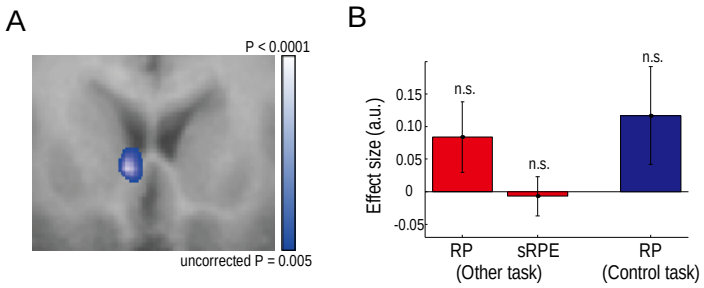


Figure S3 related to Figure 4: Reward prediction error signals in the ventral striatum (vStr) during the Control task.

(A) BOLD signals observed in the vStr reflecting the reward prediction error at the time of OUTCOME in the Control task ($P < 0.05$, corrected; Table 2; The map is thresholded at $P < 0.005$, uncorrected for display). To precisely assess striatal activity, we used a local registration procedure focusing on the anterior striatum. The normalized striatum space was first defined with reference to four landmarks (the anterior commissure and the most anterior, most dorsal, and most lateral points of the striatum), and then the functional images were transformed into that space.

(B) Effect sizes of the vStr activity (error bars= ±SEM; n=36) representing the subjects' reward probability in the Other task (RP; P=0.13, one-tailed t-test), the simulated-other's reward prediction error in the Other task (sRPE; P=0.80), and the reward probability in the Control task (RP; P=0.13). n.s. = not significant.

Supplemental Tables

Table S1 related to Figure 1. Best fitting parameter estimates

Control task		Learning rate, η	Stochasticity in the choices, β	Risk parameter γ	pseudo-R^2
RL					
	25th percentile	0.054	0.091	1.000	0.628
	Median	0.084	0.121	1.388	0.725
	75th percentile	0.099	0.202	2.700	0.786
PR					
	25th percentile	0.058	2.517	1.000	0.257
	Median	0.077	3.178	1.000	0.313
	75th percentile	0.104	4.125	1.000	0.415
MG					
	25th percentile	-	0.014	-	0.123
	Median	-	0.019	-	0.202
	75th percentile	-	0.025	-	0.255

Other task		Learning rate, η sRPE	Learning rate, η sAPE	Stochasticity in the choices, β	Risk parameter γ	pseudo-R^2
S-RL$_{sRPE+sAPE}$						
	25th percentile	0.026	0.001	0.093	0.610	0.659
	Median	0.051	0.011	0.102	1.000	0.735
	75th percentile	0.089	0.057	0.126	1.000	0.781
S-RL$_{sRPE}$						
	25th percentile	0.046	-	0.078	1.000	0.615
	Median	0.073	-	0.097	1.000	0.720
	75th percentile	0.108	-	0.105	1.000	0.752
S-RL$_{sAPE}$						
	25th percentile	-	0.014	0.073	0.529	0.569
	Median	-	0.072	0.093	0.581	0.686
	75th percentile	-	0.180	0.104	1.000	0.733
S-free RL						
	25th percentile	0.014		0.030	1.000	0.123
	Median	0.044		0.046	1.000	0.175
	75th percentile	0.063		0.064	1.000	0.230

The best-fitting parameter estimates for each model are shown as the median plus the 1st and 3rd quartiles across subjects. Also shown are medians and quartiles for the pseudo-R^2 at the best fitting parameters, a normalized measure of the degree to which the model explained the choice data (Daw et al., 2006). Abbreviations for each model are the same in Figure 1D.

Table S2 related to Figure 1. Model comparison among S-RL $_{sRPE+sAPE}$, S-RL $_{sRPE}$ and S-RL $_{sAPE}$ models

	AIC					corrected AIC			
	Mean ± SEM	Total	# Subjects favoring FS-RL	Paired *t*-test	Fit all subjects together	Mean ± SEM	Total	# Subjects favoring FS-RL	Paired *t*-test
S-RL $_{sRPE+sAPE}$	40.6 ± 2.3	1461.1	-	-	1618.8	41.0 ± 2.3	1475.2	-	-
S-RL $_{sRPE}$	43.0 ± 2.4	1547.0	20	t(35) = 3.25 P = 0.0013	1625.5	43.2 ± 2.4	1553.9	19	t(35) = 2.98 P = 0.0026
S-RL $_{sAPE}$	55.5 ± 5.5	1997.4	23	t(35) = 3.20 P = 0.0015	1778.5	55.7 ± 5.5	2005.8	23	t(35) = 3.17 P = 0.0016

	Model Evidence (negative, log)			
	Mean ± SEM	Total (GBF)	# Subjects favoring FS-RL	Exceedance Probability
S-RL $_{sRPE+sAPE}$	19.8 ± 1.3	711.2	-	1.00
S-RL $_{sRPE}$	21.2 ± 1.3	763.0	24	0.00
S-RL $_{sAPE}$	22.6 ± 1.3	813.1	28	0.00

Results of comparing the goodness of fit of the S-RL$_{sAPE}$, S-RL$_{sRPE}$, and S-RL$_{sRPE+sAPE}$ models to choice behavior in the Other task (abbreviations are the same as in Figure 1D). AIC and corrected AIC (cAIC=AIC+ $2k(k+1)/(n-k-1)$, where k and n are the number of free parameters and the sample size, respectively (Burnham and Anderson, 2002); smaller values indicate a better fit): the average and total values across subjects; the number of subjects favoring S-RL$_{sRPE+sAPE}$; and paired *t*-test over the distribution of individual subject's differences; AIC fitted to all subjects together (assuming a single set of parameters for all subjects). Bayesian model comparison based on the negative log model evidence (smaller values indicate a better fit): the average values; the total values, often called a group Bayes factor (GBF); the number of subjects favoring S-RL$_{sRPE+sAPE}$; and the Bayesian exceedance probability (Stephan et al., 2009). The so-called model evidence of each model's fit to each subject's behavior was obtained using the variational Bayes method (with factorized approximations) to integrate out the model's free parameters (Bishop, 2006); prior distributions of the parameters were assumed to be uniform (with ranges of $[0, 0.5]$ for β, $[0,1]$ for $w = \eta_{RPE}/(\eta_{RPE}+\eta_{APE})$, $[-7,15]$ for $\log(\gamma)$). To compute the exceedance probabilities, we used the spm_BMS routine from SPM8 (http://www.fil.ion.ucl.ac.uk/spm/software/spm8/).

Supplemental Experimental Procedures

Subjects

Thirty-nine healthy, normal subjects (11 females, 28 males; age range: 20-35 years; mean ± standard deviation, 22.6 ± 4.0) participated in the fMRI experiment. Subjects were pre-assessed to exclude those with any previous history of neurological or psychiatric illness. Before the experiment, subjects were instructed about the experimental tasks, and informed that they would receive monetary rewards proportional to the average of all the points they earned in four test sessions (two fMRI scan sessions, from which the results of both behavioral and imaging data are reported in the main text, and two other sessions not involving fMRI, the results of which were not reported in the main text; see below) in addition to a base participation fee (6000 yen). The total monetary reimbursement in Yen equaled 200 □ (average points – 20) + 6000. A separate behavioral experiment (see Figure 1C) involved 24 normal subjects (11 females, 13 males; age range, 18-24 years; mean, 20.0 ± 1.2 years) who did not participate in the fMRI experiment. The procedures used were virtually identical to those used in the fMRI experiment. These subjects received monetary rewards based on the points they earned during three experimental sessions (see below) in addition to the base fee. All subjects in both experiments gave their informed written consent, and the study was approved by RIKEN'S Third Research Ethics Committee.

Experimental tasks

Two tasks, the *Control* and the *Other*, were conducted (Figure 1A). Each task consisted of multiple trials in which different pairs of fractal stimuli were used. Each trial within both tasks consisted of four phases.

The Control task was a one-armed bandit task (Behrens et al., 2007), in which subjects were instructed to choose the stimulus that would maximize the number of points earned. At the beginning of each trial, subjects were presented with a pair of fractal stimuli with a fixation point between them. The two stimuli with randomly assigned reward magnitudes, indicated by numbers in their centers, were randomly positioned left or right of the fixation point in every trial (for 3-7 s; CUE phase; Figure 1A). In every trial, the reward magnitude for one stimulus (R) was randomly sampled from a uniform distribution ranging from 1 to 99 points, while the reward for the other stimulus was set to (100-R); this randomization was further constrained to ensure that the same stimulus was not assigned the higher magnitude in three successive trials. This constraint was introduced, in addition to reward magnitude randomization, to further ensure that subjects did not repeatedly choose the same stimulus (see the control analysis described below). When the fixation point was changed to a question mark, subjects made their choice by pressing a button with their right hand within 1.5 s (RESPONSE phase). The chosen stimulus was immediately highlighted by a gray frame, initiating the INTER-STIMULUS INTERVAL (ISI) phase. After the ISI phase (3-7 s), the rewarded stimulus was revealed in the center of the screen for 3 s (OUTCOME phase). This was followed by a 3-5 s intertrial interval (ITI) before the next trial was commenced. In both the Control and Other tasks, one of the two stimuli was arbitrarily designated to have a higher reward probability (set to be 0.75, and thereby setting the other stimulus probability to 0.25). Subjects were not informed of the probability, but were instructed that the reward probabilities were independent of the reward magnitudes.

In the Other task, subjects were instructed to predict the choice of another person who had performed the Control task. From the CUE to the ISI phase, the images on the screen were identical to those in the Control task in terms of presentation. However, the two stimuli

presented in the CUE phase were generated for the other person performing the Control task. Upon appearance of the question mark at the fixation point (RESPONSE), subjects predicted the choice made by the other person; this choice was immediately highlighted by a gray frame, initiating the ISI. In the OUTCOME phase, the other person's actual choice was highlighted by a red frame, and the rewarded stimulus for the other was indicated in the center. For every trial in which the subjects' predicted choice matched the other's actual choice, they earned a fixed reward of 50 points. The Other task was designed to minimize differences from the Control task, so that the number of phases was the same between the two tasks and in terms of visual presentation, only the red frame, indicating the other's choices, was added at the OUTCOME phase in the Other task.

Subjects were told they would see on the screen the choices of another subject who had participated in previous experiments. However, the choices of the other subject were actually generated by an RL model (see below). In the Other task in the fMRI experiment, the RL model generated choices on a risk-neutral basis; the model's parameters ($\eta=0.14, \beta=0.098, \gamma=1$; see below) were determined from average values obtained in a pilot experiment (independent from the experiments reported in this study). Accordingly, the choices generated by the model were considered to approximately mimic average (risk-neutral) human behavior in this task, and thus allowed us to use the same type of the other's behavior for all subjects; this approach was supported by a separate behavioral control analysis (see below). In post-experiment interviews, we debriefed each subject and confirmed that they had no doubt that the choices were being made by someone else.

For the experiment in the MRI scanner, two tasks, one Control and one Other, were employed. Each task consisted of 90 trials, and the order of the two tasks was counter-balanced across subjects. Before these tasks, subjects performed a short exercise session (Control task, 20

trials) inside the MRI scanner. Before entering the scanner, they were first familiarized with the tasks through performance of a few tens of trials in both tasks using a shorter timing sequence for the phases, after which they performed two test sessions: 120 trials of both the Control and Other tasks, the order of which was counter-balanced across subjects. Subjects also obtained earnings in both test sessions. The results of these pre-scanning sessions were not reported in the present paper, as they were essentially the same as those from the two fMRI sessions reported in the paper. Finally, subjects performed the two tasks (40 trials for each) with the same timing sequence used for experiments involving the MRI scanner.

Three conditions were used in a separate behavioral experiment (Figure 1C): one Control and two Others, and the order of the three was randomized across subjects. As in the fMRI experiment, these additional subjects also went through a training session before starting the main experiment. The settings for the Control and 'Other I' task were the same as described for the fMRI experiment, but in the 'Other II' task, a risk-aversive RL model ($\eta = 0.14, \beta = 0.098, \gamma = 1.568$) was used to generate the other's choices instead of the risk-neutral model. After altering the magnitude of γ while fixing the magnitudes of the other two parameters, as in the original Other task, the RL model was found to choose the stimulus with the higher reward probability in the Control task with the same average percentage as the subjects in the fMRI experiment who behaved risk-aversively – i.e., there was no statistical difference after 100 runs of the model. After completing the experiments, we asked subjects via questionnaires: (*i*) Which information did you use for predicting the other's choices in the Other task: the other's outcomes, the other's choices, or both?; (*ii*) Did you notice any differences between the other's behaviors under the two different Other conditions? A majority of subjects reported that (i) they considered both sources of information (20/22 subjects) and (ii) they noticed the difference in the two conditions (21/22 subjects). These answers are further evidence

that subjects simulated the other's value-based decision making and used both the simulated-other's reward error and action prediction error.

Behavioral analysis and computational models fitted to behavior

Among the 39 subjects who underwent fMRI scans, three were not included in the final analyses because their choice behaviors were found to be outliers in the pool of subjects ($P < 0.01$, Thompson's test). The remaining 36 subjects were used for the subsequent behavioral and fMRI data analyses. For the behavioral analyses shown in Figure 1C, two of the 24 subjects were not included due to outlier behavior ($P < 0.01$, Thompson's test), leaving 22 subjects for the final analysis.

We fitted several computational models to the subjects' choice behaviors in both tasks. All of these models were based on and modified from the Q learning model, a basic RL model (Sutton and Barto, 1998), which is referred to simply as the RL model, hereafter (Supplemental Figure S1A). In the Control task, the RL model, being risk-neutral, constructed values Q_s of the two stimuli in each trial, given by

$$Q_S(A) = R_S(A) \cdot p_S(A), \quad (1)$$

where $R_S(A)$ is the reward magnitude of stimulus A in a given trial, $p_S(A)$ is the reward probability of stimulus A, and the subscript, *s*, refers to the subject (under simulation-free RL formulation). The value of the other stimulus, B, was similarly derived; for simplicity, therefore, we will only provide equations for stimulus A. To account for possible risk-aversive (or risk-prone) behaviors of subjects, we followed the approach taken by Behrens et al. (2007); we included a free parameter that replaced $p_S(A)$ in Eq (1) with $F(p_S(A),\gamma)$; where γ is a non-negative free parameter for risk behavior, and the function $F(p,\gamma)$ is a simple non-linear

transform within the bounds of 0 and 1, given by

$$F(p,\gamma) = \max\left[\min\left[\left(\gamma(p-0.5)+0.5\right),1\right],0\right]. \quad (2)$$

When $\gamma = 1$, $F(p_S(A),\gamma) = p_S(A)$, leading to risk-neutral behavior, whereas $\gamma > 1$ and $\gamma < 1$ imply risk-aversive and risk-prone behavior, respectively.

The RL model chose either stimulus A or B based on the choice probability (of stimulus A) $q_S(A)$, given by

$$q_S(A) = f\left(Q_S(A) - Q_S(B)\right), \quad (3)$$

where $f(z) = 1/\left[1+\exp\{-\beta z\}\right]$ is a sigmoidal function allowing probabilistic choices with a free parameter β, which adjusts the degree of stochasticity in the choices (Sutton and Barto, 1998). Once a choice was made and the reward outcome was revealed, the RL model utilized the reward prediction error to update the stimulus value based on the Rescorla-Wagner rule. In the context of our tasks, only the reward probability was updated (Behrens et al., 2007) because this was the only variable unknown to subjects. Accordingly, when stimulus A was chosen, the reward prediction error was given by

$$\delta_S = r_S - p_S(A), \quad (4)$$

where r_s is the reward outcome (1 if stimulus A is rewarded and 0 otherwise). The reward probability was updated using $p_S(A) \leftarrow p_S(A) + \eta\delta_S$, where η, another free parameter, is the learning rate.

Two variants of the RL model were also fitted to the behavior in the Control task. To compute the stimulus values, the two models ignored either the reward magnitude or the reward probability, thus setting $Q_S(A) = p_S(A)$ or $Q_S(A) = R_S(A)$, respectively. We also tested a

model using $Q_S(A) = F(p_S(A), \gamma)$, but its fit was significantly worse than that of the RL model (data not shown).

The model with the best fit to the behavior in the Other task was the model that we called Simulation-$RL_{sRPE+sAPE}$ model (S-$RL_{sRPE+sAPE}$) (see Supplemental Figure S1A). In each trial, the S-$RL_{sRPE+sAPE}$ model computed the subject's choice probability $q_{\tilde{s}}(A) = f(Q_{\tilde{s}}(A) - Q_{\tilde{s}}(B))$, where we used $\tilde{s}$ instead of s to indicate subjects, because $q_{\tilde{s}}(A)$ was computed in a "simulation-based" manner – i.e., by simulating the other's RL model. We reserved $\boldsymbol{s}$ to indicate subjects when computing in a "simulation-free" manner. Here, $Q_{\tilde{s}}(A) = R_S \cdot p_{\tilde{s}}(A)$ indicates stimulus A's value for subjects; R_S $(= R_S(A) = R_S(B))$ denotes the fixed reward outcome that subjects would obtain if their prediction of the other's choice matched the other's actual choice. When simulating the other's RL model, the subjects' stimulus reward probability is equivalent to the simulated-other's choice probability, $p_{\tilde{s}}(A) = q_O(A)$. The simulated-other's choice probability as well as the simulated-other's value of stimulus A are given by

$$q_O(A) = f(Q_O(A) - Q_O(B)) \text{ and } Q_O(A) = R_O(A) \cdot p_O(A), \quad (5)$$

where $R_O(A)$ is the reward magnitude of stimulus A for the other in the trial, and $p_O(A)$ is the simulated-other's reward probability for stimulus A. When inclusion of the risk parameter produced a better fit to behavior, $p_O(A)$ in the second equation was replaced by $F(p_O(A), \gamma)$.

When the outcome for the other was revealed (denoted by r_O, which was 1 if the other received a reward and 0 otherwise), the S-$RL_{sRPE+sAPE}$ model updated the reward

probability, not only using the simulated-other's reward prediction error but also the simulated-other's action prediction error. The simulated-other's reward prediction error was given by $\delta_O(A) = r_O - p_O(A)$. The simulated-other's action prediction error was generated first in the 'action' level as the difference between the other's actual choice and the simulated-other's choice probability, given by $\sigma'_O(A) = I_A(A) - q_O(A) = 1 - q_O(A)$, wherein the choice probability is generated through a sigmoid function using the difference of two values (1st equation in Eq (5)); thus, to be used for updating the simulated-other's value, the action prediction error needed to be 'pulled back', or transformed, from the action to the value level (Supplemental Figure S1A). As this error should act as a learning signal to update the simulated-other's value, which is to cause a small change of the value in the value level, the transformation of the error between the two levels can be formulated by making correspondingly small changes in both of the levels. This is accomplished using a general notion of variation. Given function $z = f(x)$, a variation equation is given by $\delta z = \partial f(x)\delta x$, which indicates how small changes between both sides ($\delta z, \delta x$) should match, and in our case, z and x correspond to the simulated-other's choice probability and the chosen value, respectively. Applying the variation formulation to our case leads to,

$$\delta q_O(A) = \left[\partial f\left(Q_O(A) - Q_O(B)\right) / \partial Q_O(A)\right]\delta Q_O(A). \quad (6)$$

When we set $\delta q_O(A) = \sigma'_O(A)$ and replaced the 1st term on the left hand side of the equation with K, we let $\delta Q_O(A) = \delta q_O(A)/K$ when $K \neq 0$; otherwise $\delta Q_O(A) = 0$. By simple calculation, we obtain $K = R_O(A) q_O(A) q_O(B)$, where β is omitted on the right side because it will be absorbed into the learning rate. Thus, the simulated-other's action prediction

error (in the value level) is given by $\sigma_O(A)=\sigma'_O(A)/K \quad (K\neq 0)$; we refer to the simulated-other's action prediction error as being in the value level, unless explicitly stated otherwise. Then, the S-RL$_{sRPE+sAPE}$ updated the simulated-other's reward probability, using both the simulated-other's reward and action prediction errors together, given by

$$p_O(A) \leftarrow p_O(A)+\eta_{sRPE}\delta_O(A)+\eta_{sAPE}\sigma_O(A), \quad (7)$$

where the two η's indicate the learning rates of the reward and action prediction errors. In both the Control and Other tasks, the learned variable is a reward probability dissociated from reward magnitudes (Behrens et al., 2008; Behrens et al., 2007; Boorman et al., 2009), as magnitudes were randomly assigned to the stimuli, and independent of the stimulus (see below for the control analysis confirming this view).

The two other Simulation-RL models, each using only one of the two prediction errors, were modeled by using either $\eta_{sRPE}\delta_O$ or $\eta_{sAPE}\sigma_O$ to update $p_O(A)$ in Eq (7). The simulation-free RL model, which focused only on the subjects' own outcomes during the Other task, set the choice probability to $q_S(A)=f(Q_S(A)-Q_S(B))$, where $Q_S(A)=R_S\cdot p_S(A)$, given the subjects' reward R_S and the estimated reward probability $p_S(A)$. $p_S(A)$ was replaced by $F(p_S(A),\gamma)$ whenever a better fit to behavior was obtained by including the risk parameter. The reward probability was updated by $p_S(A)\leftarrow p_S(A)+\eta\delta_S$ using the reward prediction error $\delta_S(C_S)=r_S-p_S(A)$.

We used a maximum likelihood approach to fit the models to the subjects' behaviors. For individual subjects, we minimized the negative log-likelihood of the sum of each model's choice probabilities against the actual choices made by subjects (*matlab* command *fminsearch*;

Matlab R2007b, MathWorks). Each minimization was repeated 50 times, using randomly generated initial values. The model with the best minimization was then selected, which also determined the estimated values of the model's free parameters. For comparisons of goodness of fit, we used Akaike's Information Criterion (AIC) to take into account the different numbers of free parameters between models. We first compared the total AIC values between two models, calculating each AIC value as a summation of all subjects' AIC values. Second, to take into account variation in the AIC values across subjects, we also used a paired *t*-test to compare the distribution of differences in the AIC values obtained in the two models. When the results of the two comparisons were consistent, we reported the results of the second analysis in the Results (e.g., in Figure 1D), since this was more stringent; otherwise, we reported the results for both comparisons. For a given model's fit to each subject's behavior in a task, the inclusion of the risk parameter was determined using the AIC value to compare the fit by two variants of the given model, with or without including the risk parameter (the risk parameter, when included, was optimized together with the other parameters in the minimization); the risk parameter was included only if it yielded a better fit for the given model with the subject in the task.

When we reported the accuracy of each model's performance averaged across subjects in Results, the accuracy was expressed as a percentage, across trials within a given task, of the model's stimuli with the higher choice probability that matched the stimulus actually chosen by subjects.

Given that the S-$RL_{sRPE+sAPE}$ model had the best fit to the behavior in the Other task, we performed two control analyses, which provided evidence supporting separate contributions of the simulated-other's reward and action prediction errors to simulation learning. First, we examined Spearman's correlation coefficient between the two errors; it was found to be low across subjects (mean ± standard deviation: -0.018 ± 0.129); at each individual, only 2 of 36

subjects had correlations significantly different from zero ($P < 0.05$, two-tailed; the two subjects' correlation coefficients corresponded to the maximum and minimum magnitudes among the subjects, 0.198 and -0.271, respectively). Also, the correlation between the learning rates of the two errors was low (-0.155), insignificantly different from zero (P=0.366). As a further confirmation, we also examined Spearman's correlation between the information provided by the other's rewards and actions (using a binary representation); it was low (-0.042 ± 0.125); at each individual, only 2 subjects (who were different from the two subjects above) had correlations significantly different from zero ($P < 0.05$, two-tailed; the two subjects' correlation coefficients corresponded to the maximum and minimum, 0.300 and -0.277, respectively). Together, these results indicate that the two errors can in principle have a separate contribution to learning to simulate the other's decisions.

Second, we conducted a two-way repeated measures ANOVA analysis to examine whether the subjects' behavior differs with respect to the magnitudes of the simulated-other's action prediction error in the previous trial (Supplemental Figure S1B). The $\text{S-RL}_{\text{sRPE+sAPE}}$ differs from the $\text{S-RL}_{\text{sRPE}}$ in that it uses the action prediction error as an additional learning signal. Therefore, the $\text{S-RL}_{\text{sRPE+sAPE}}$ should predict that, in a given trial, the subjects are more inclined to choose the same option that the other chose in the previous trial, as the simulated-other's action prediction error is larger; whereas the $\text{S-RL}_{\text{sRPE}}$ is insensitive to the information of this error. We examined this hypothesis by analyzing the percentage of times that, in a given trial, the subject's choice coincided with the other's chosen option in the previous trial. For the first variable of the ANOVA, we used the median of the action prediction error (for each subject) to sort the trials during the Other task into two groups. As the simulated-other's reward prediction error also contributes to learning, it was also of particular interest to contrast the cases when the reward prediction error was either negative or positive, because the effects of

the reward and action prediction errors on updating the simulated-other's reward probability are opposite only when the reward prediction error is negative. Thus, as the second variable in the ANOVA, we used the sign of the reward prediction error to also classify the trials into the two groups.

We also examined the fit of several variants of the S-RL$_{sRPE+sAPE}$ model to the behavior in the Other task, compared with that of the original S-RL$_{sRPE+sAPE}$ model. First, we examined two variants including risk parameters differently from the original model and the results of the comparison of the fit indicated that the original S-RL$_{sRPE+sAPE}$ model fit equally or better to the behavior compared with the two variants. In the original model, the risk parameter was included in the simulated-other's choice probability, but not in the subject's own choice probability. This was because we reasoned that the effect of the risk parameter was relatively negligible at the subject's level of valuation, as the reward magnitude was fixed for subjects. However, we also examined two other variants of the S-RL$_{sRPE+sAPE}$ model: one that included a risk parameter only at the subject's level and another that included risk parameters at both the subject's and simulated-other's levels. The original S-RL$_{sRPE+sAPE}$ model fit the behavior significantly better than the variant that included a risk parameter only at the subject's level (1461.1 vs. 1543.1; in total AIC values and $P < 0.01$ by paired *t*-test). The original was also significantly better than the variant that included risk parameters at both levels (1461.1 vs. 1466.4; total AIC values), though the original did not significantly differ from the variant based on a paired *t*-test ($P = 0.36$).

Second, to examine a variant of the S-RL$_{sRPE+sAPE}$ model that used the simulated-other's action prediction error only for biasing the subject's choices in the next trial, $\eta_{sAPE}\sigma_O(A)$ was omitted from Eq (7) and the subject's choice probability was modified as

$q_{\tilde{S}}(A) = f\left(Q_{\tilde{S}}(A) - Q_{\tilde{S}}(B)\right) + \alpha\sigma_O^{'}(A)$, where α is a free parameter to determine the influence of the bias (determined when maximizing likelihood), and $\sigma_O^{'}$ is the simulated-other's action prediction error in the action level from the previous trial (Note: We also examined the case in which we the action prediction error in the value level was use for biasing, i.e. using $q_{\tilde{S}}(A) = f\left(Q_{\tilde{S}}(A) - Q_{\tilde{S}}(B)\right) + \alpha\sigma_O(A)$; the result was the same: $P <$ 0.001, one-tailed paired *t*-test).

Additional control analyses on behavior and computational models fitted to behavior

In addition to the results reported in the main text, we summarize here further control analyses using the same data in the main text.

We conducted a control analysis on the non-linear risk function (Eq. 2) for capturing the subjects' risk tendency observed in experimental tasks. Overall, the results of the control analysis support the use of the non-linear risk function; or at the very least, there is no reason to believe that the other functions examined in the control analysis can better account for the risk behavior. We examined two other representative approaches accounting for risk behavior: using the power function or mean-variance functions, both of which are often used in neuroscience studies (Huettel et al., 2006; Tobler et al., 2009). The power function had the form, $Q(A) = \{R(A)\}^{-\gamma} \cdot p(A)$, where subscripts were dropped for simplicity and γ in the power of the reward magnitude is the risk parameter in this function. The mean-variance function is,

$$Q(A) = E[R(A)] - \gamma \text{Variance}[R(A)] = R(A) \cdot P(A) - \gamma\{R(A)\}^2 \cdot P(A)(1 - P(A)),$$

where again γ is the risk parameter of the function. First, we compared the fits of the three functions to the subjects' behavior in the Control task. The non-linear function (Eq. 2) fit the

behavior equally as well as the power and mean-variance functions (P = 0.12 and P = 0.18, respectively, by paired t-test over the AIC distributions). The correlations of the risk parameter values between the non-linear function and each of the two other functions is very high (Spearman's correlation coefficient: 0.93 and 0.95 for the power and mean-variance function, respectively), suggesting that they capture the nature of the risk behavior in a very similar way. Second, we then examined a variant of the power function, given by $Q(A) = \{R(A)\}^{-\gamma} \cdot \{p(A)\}^{-\gamma'}$; that was used in another study (Boorman et al., 2009), in which a task was somewhat similar to the Control task in this study. This model's fit was again not different from that used in the main study (P = 0.37). Third, we also compared these different approaches for fitting the models to the behavior in the Other task; the non-linear function had a comparable or better fit than the other three functions ($P < 0.05$ with the original power function, $P < 0.01$ with the mean-variance function, and P = 0.18 with the variant of the power function).

We conducted a control analysis to address a possible concern of whether reward magnitudes might have an effect on learning the reward probability; for instance, missing out on a large reward magnitude might have a particular effect on learning, compared to missing out on a relatively small reward. In the original model settings for both Control and Other tasks, we did not include any parameters that might take into account effects of reward magnitudes on learning reward probability. This was because in our experimental tasks, reward magnitudes were randomized every trial, independently (or almost completely independently at the very least given the additional constraint on reward magnitude randomization) from the reward probability of the stimulus. Thus we consider it neither possible nor necessary to learn to associate specific reward magnitudes with specific stimuli, as supported by earlier studies using the same or similar task for the Control task (Behrens et al., 2007; Boorman et al., 2009). Nevertheless, to address the concern, we examined the behavioral fits of several variants of the

models best fitted to each task (the RL model in the Control task and S-RL$_{sRPE+sAPE}$ model in the Other task). Using two approaches, we constructed models' variants that had different learning parameters depending on reward magnitudes as well as on whether reward was gained or missed. The first approach was to allow different learning rates for when the reward magnitude of the stimulus chosen in the trial was smaller or larger than 50 (since reward magnitudes were randomly assigned to the two stimuli as *R* and 100-*R*). Thus, the variant of the RL model in the Control task (hereafter called the 1st-variant RL model) had two learning rate parameters, only one of which was used for updating the value, depending on the magnitude of the stimulus chosen by the subject in the trial. Similarly, for the S-RL$_{sRPE+sAPE}$ model in the Other task, we allowed different learning parameters depending on the reward magnitude (for the other) of the stimulus chosen by the other in Other task. There were three variants of the S-RL$_{sRPE+sAPE}$ model; the '1st-R-variant' and '1st-A-variant' S-RL$_{sRPE+sAPE}$ model had the two different learning parameters only for the simulated-other's reward and action prediction error, respectively; and the '1st-RA-variant' S-RL$_{sRPE+sAPE}$ model had the two different learning parameters for each of the two errors. The second approach was to further allow different learning parameters depending on whether the reward was obtained or missed. Thus, this variant of the RL model in the Control task (the 2nd-variant RL model) had four learning parameters, one of which was used in each trial, depending on the magnitude of the stimulus chosen by the subject in the trial and on whether the chosen stimulus was rewarded or not. For the S-RL$_{sRPE+sAPE}$ model, there were three-variants; the '2nd-R-variant' and '2nd-A-variant' of the S-RL$_{sRPE+sAPE}$ model had the four different learning parameters only for the simulated-other's reward and action prediction error, respectively, and the '2nd-RA-variant' S-RL$_{sRPE+sAPE}$ model had the four different learning parameters for each of the two errors.

The comparison of the fit of these variants to the behavior with that of the original model

(using the paired *t*-test over the AIC distributions) demonstrated that reward magnitudes did not have a noticeable effect on learning the reward probability; For the Control task, the fit of the original RL model was not significantly different from those of the two variants (the 1st-variant and the 2nd-variant RL model; P =0.15, and P =0.61, respectively); For the Other task, the fit of the original S-RL$_{sRPE+sAPE}$ model was not significantly different from those of most of the variants (the 1st-R-variant, the 1st-A-variant, the 1st-RA-variant, the 2nd-R-variant and the 2nd-A-variant S-RL$_{sRPE+sAPE}$ model; P =0.22, P =0.12, P =0.10, P =0.42, and P =0.63, respectively) and was better than that of the most complex variant (the 2nd-RA-variant S-RL$_{sRPE+sAPE}$, $P < 0.01$).

Additional behavioral experiments for control analyses

In addition to the results reported in the main text, we further performed two additional behavioral experiments for control analyses that are summarized here. Subjects in each experiment did not participate in any other experiments described in this report, and received monetary rewards, in addition to the base fee, based on the points they earned during the experiment. The procedures used were virtually identical to those used in the fMRI experiment except particular aspects of the experiment (described below). After completing the experiment, we asked subjects to fill in the questionnaires. All subjects gave their informed written consent, and both studies were approved by RIKEN'S Third Research Ethics Committee.

In the first experiment, we examined the question of whether the subjects' prediction of the other's choices generated by a computer model (which was adopted in the main study) may differ from those made predicting the choices generated by actual humans, or more precisely, by risk-neutral humans. An additional question was whether the subjects' predictions were actually meaningful or at all different from those made when the other's choices were random choices,

i.e., generated by a random-chooser. This additional experiment involved 17 normal subjects (11 females, 6 males; age range, 18-33 years; mean, 21.4 ± 4.0 years). The subjects earned the points during the four experimental sessions, corresponding to the following four conditions: one Control task and three Other tasks. The other's choices were generated by the RL model (O-RL, using the same parameter values used in the main report), by a risk-neutral human (O-human), and by a random-chooser (O-random). The procedures were modified from those used in the fMRI experiment in the following two aspects: (i) O-random was not used in exercise sessions and was always placed in the last of the four main sessions (the order of the other three sessions was counter-balanced across subjects). This was to avoid any potential compounds. When we compared the subjects' behaviors in the O-RL and O-random tasks prior to this experiment, they were already quite different; thus, if the subjects had experienced O-random either in exercise sessions or as one of the main sessions before the other main sessions, they might have been confused by the experience, which might have affected their behavior in the other main sessions. (ii) We balanced the subjects' experience of O-RL and O-human tasks in the exercise session. For the short exercise session, either O-RL or O-human was used to generate the other's choices, counter-balanced across the subjects. For the long exercise session, the subjects experienced three sessions: one Control task and two Other tasks (O-RL and O-human). The choices in the O-human task were those of actual human subjects during the Control task, who indicated risk-neutral behavior in the behavioral experiment reported in the main text (i.e., the experiment conducted for the results in Figure 1C); there were 8 subjects in this pool. For each subject in this experiment, a different set of O-human choices was randomly chosen in the exercise session; the sets of choices were also randomly chosen in the main session but we ensured that they were different from that used in the exercise session. Among the 17 subjects, there were no outliers ($P > 0.01$, Thompson's test) and all data was

included in the subsequent analysis.

The results of the additional experiment support the rationale for the use of the fitted risk-neutral RL model in the main report (see Supplemental Figure S1C and the legend). This conclusion was further supported by the subjects' answers to the following post-experiment questionnaires**:** (*i*) Which information did you use for predicting the other's choices in the Other task: the other's outcomes, the other's choices, or both? (*ii*) Did you notice any differences in the other's behavior among the three Other task sessions; if yes, which session(s) were different from the other sessions? (*iii*) Were there any of the three Other task sessions that you felt that the other behaved non-humanly? A majority of subjects reported that (*i*) they considered both sources of information (14/17 subjects), (*ii*) they considered that the O-random (i.e., the last of the three Other task sessions) behaved differently from the O-RL and O-human (14/17 subjects), and (*iii*) they considered that the O-random behaved non-humanly (13/17 subjects).

In the second experiment, we examined whether our additional constraint on the reward magnitude randomization (such that the same stimulus was not assigned the higher magnitude in three successive trials) might alter the subjects' behavior, compared to the case when the reward magnitude assignment was completely random. This additional experiment involved 23 normal subjects (9 females, 14 males; age range, 18-37 years; mean, 20.9 ± 4.2 years). The subjects earned the points during the following four experimental sessions: the Control and Other tasks when the reward magnitude randomization was conducted with or without the constraint mentioned above. The procedures used were modified for the following; we tried to ensure that the subjects experienced the tasks equally with or without the constraint during exercise sessions before the main sessions. For the short exercise session, the subjects experienced both the Control and Other tasks either with or without the constraint, counter-balanced across subjects, while during the long exercise session, they experienced all the four conditions. After

excluding two subjects based on their outlier choice behaviors ($P < 0.01$, Thompson's test), the remaining 21 subjects (9 females, 11 males; age range, 18-37 years; mean, 20.8 ± 4.4 years) were used for the subsequent analysis.

The results of this additional experiment demonstrate that the subjects' behaviors in both the Control and Other tasks did not significantly differ with or without the additional constraint on reward magnitude randomization (see Supplemental Figure S1D and the legend). This conclusion was further supported by the subjects' answers to the following post-experiment questionnaires: (*i*) Which information did you use for predicting the other's choices in the Other task: the other's outcomes, the other's choices, or both? (*ii*) Did you notice any differences in the reward magnitudes of the various options or in the 'correct' stimulus between the two sessions of the Control task? (*iii*) Did you notice any differences in the reward magnitudes of the options or in the 'correct' stimulus between the two sessions of the Other task? A majority of subjects reported that (i) they considered both sources of information (18/21 subjects), (ii) they noticed no differences in the two sessions of the Control task (19/21 subjects), and (iii) they noticed no differences in the two sessions of the Other task (20/21 subjects).

fMRI acquisition and analysis.

The fMRI images were collected using a 4 T Varian Unity Inova MRI system (Agilient Inc., Santa Clara, CA) with a phased array coil (four receiver coils were placed over the left and right frontal and occipital cortices). For subjects positioned in the scanner, visual input was provided via a fiber optic goggle system (Avotec, Jensen Beach, FL) that subtended 25°×19° of the visual angle, and subjects used a button box to make their responses. The BOLD signal was measured using a two shot T2*-weighted echo planar imaging sequence (Volume TR=2222 ms, TE=20.5 ms, FA=30°). Twenty-five axial slices (thickness=3.0 mm, gap=1 mm,

FOV=192×192 mm, matrix=64×64) parallel to the AC-PC plane were acquired per volume. The start of an experimental task was synchronized with the first EPI acquisition timing. Before, after, or between the functional runs, a set of high-resolution (1 mm^3) and a set of low-resolution (1.72 mm^3) whole-brain anatomical images were acquired using a T1-weighted 3D FLASH pulse sequence (TI=500ms, FA=11°, [TR=12.7ms, TE=6.8ms] for the high resolution scans, [TR=11.1 ms, TE=6.2 ms] for the low resolution scans). The low-resolution anatomical imaging slices were parallel to the functional imaging slices and were used to aid in co-registering the functional data to the high-resolution anatomical data. A pressure sensor was used to monitor and measure the respiration signal, and a pulse oximeter was used to measure the cardiac signal. The respiratory and cardiac signals were used in postprocessing to remove physiological fluctuations from functional images (Hu et al., 1995).

Functional and anatomical images were analyzed using Brain Voyager QX 2.1 (Brain Innovation B.V., Maastricht, NL). Functional images for each subject were preprocessed, which included slice time correction, three-dimensional motion correction, spatial smoothing with a Gaussian filter (FWHM=8 mm), and high-pass filtering (three cycles per run length). Anatomical images of each subject were transformed into the standard Talairach space (TAL) (Talairach and Tournoux, 1988). Functional images were then normalized and resized according to transformed structural images, and thus transformed into the standard Talairach space. An exception was activation in the ventral striatum reported in Supplemental Figure S3 (see legend).

We employed a so-called model-based analysis (O'Doherty et al., 2007) to analyze the BOLD signals in both tasks. For the Control task, we created subject-specific design matrices containing the following regressors: (1) six regressors encoding the average BOLD responses for the onsets and the periods of the DECISION, ISI, and OUTCOME phases, where the

DECISION phase was defined as the period from the onset of CUE until subjects made their responses in the RESPONSE period and the other two phases were defined as in Figure 1A; (2) two regressors for the two variables of interest: one representing the reward probability (RP) of the stimulus chosen in the DECISION period and the other representing the reward prediction error (RPE) in the OUTCOME period. For the Other task, subject-specific design matrices contained the following regressors: (1) the same six regressors as in (1) above in the Control task; (2) three regressors for the three variables of interest: one representing the subject's reward probability (RP) for the stimulus chosen in the DECISION period, and the other two representing the simulated-other's reward (sRPE) and action prediction (sAPE) errors in the OUTCOME period. For both tasks, all regressors were convolved using a canonical hemodynamic response function. Also included were six variables of no interest – i.e., motion correction parameters – to account for motion effects. Together, these regressors were fitted to each subject's data individually, and the fitted coefficient values of the regressors (effect sizes) were then entered into a random-effects analysis and analyzed using a one-tailed *t*-test. The significances of the BOLD signals were reported based on corrected *p*-values ($P < 0.05$), using a family-wise error for multiple comparison corrections, where cluster-level inference was used. We first thresholded contrast maps at $P < 0.005$ (uncorrected) and determined the appropriate spatial extent threshold for corrected cluster-level inference at $P < 0.05$ (corrected), referring to the AlphaSim program in Analysis of Functional NeuroImages (AFNI) (Cox, 1996); This resulted in reporting uncorrected $P < 0.005$ and cluster size > 56 unless otherwise explicitly stated.

Additional regression analyses were employed to further examine the potential confounders of the variables of interest. For each variable of interest, additional regressors corresponding to potential confounders for that variable were added to the same phase of the

original regression matrix in each task, as described in the Results section. All of the signals in the vmPFC, dmPFC, and dlPFC reported in the Results section remained significant ($P < 0.05$, corrected) with these additional regressions.

To extract cross-validated percent changes in BOLD signals (Figure 2B, D), we followed the previously described leave-one-out procedure (Gläscher et al., 2010) to provide an independent criterion for ROI selection and thus ensure statistical validity (Kriegeskorte et al., 2009). We re-estimated our second-level analysis 36 times, always leaving out one subject. Starting at the peak voxels for the focal signal (e.g., the simulated-other's reward prediction error in the vmPFC in Figure 2B), we selected the nearest maximum in these cross-validation second-level analyses. The selected voxel was defined as an ROI, and we extracted the BOLD signal in the ROI from the left-out subject. Based on the magnitude of the focal signal, the left-out subject's cross-validated BOLD changes were binned as low, medium, or high (corresponding to the 33rd, 66th, and 100th percentiles, respectively) to obtain the individual's bin-wise mean BOLD changes. Then the mean BOLD changes across subjects (and the SEM) were computed for each bin. To determine whether the BOLD changes increased with the order of the bins, we calculated Spearman's correlation coefficient (ρ) using the distributions of the individual's bin-wise values, which were the difference between the individual bin-wise mean and the individual's grand mean for all the trials. Its statistical significance was tested using a one-tailed *t*-test.

To investigate the correlations between the variabilities of the subjects' effect sizes in respective brain regions and their behavioral variabilities (Figure 3), we calculated Spearman's correlation coefficient (ρ) and tested its statistical significance using a one-tailed *t*-test. Given our hypothesis that the neural variability in a ROI for each error should be positively correlated with the behavioral variability, we also examined the bootstrap test, allowing replacements and

generating 10,000 bootstrapped datasets, to examine the significance (these results are not shown, as the results are the same as those from the one-tailed *t*-test). We chose to use the Spearman's, because it is nonparametric and thus known as being relatively robust against possible outliers. Nevertheless, we performed two additional correlation analyses, each of which was more robust against possible outliners. One was to use Jackknife to detect potential outliers and remove the detected data points before computing the correlation (ρ) (Efron, 1992); in brief, we first computed the so-called Jackknife influence function $u_i\{\rho\} = (n-1)(\rho_{()} - \rho_{-i})$, where n is the number of samples, ρ_{-i} is the Spearman's correlation coefficient computed by excluding the *i*-th subject, and $\rho_{()} = \frac{1}{n}\sum_{i=1}^{n}\rho_{-i}$. We then obtained the relative Jackknife influence function, $u_i^{\dagger}\{\rho\} = u_i\{\rho\} / \sqrt{Z}$, where Z is a normalization factor given by $Z = \frac{1}{n-1}\sum_{j}^{n}(u_j\{\rho\})^2$. Using the values of the relative Jackknife influence function, we detected samples that might have rather extraordinary influences on the statistics of interest (correlation in our case). We classified these samples as outliers, if the *i*-th sample had $|u_i^{\dagger}\{\rho\}| \geq 2$. After removing the outliers, the Spearman's correlation coefficient was computed and the significance was examined using a one-tailed *t*-test. The other was to use the so-called robust correlation coefficient (Abdullah, 1990), instead of Spearman's, as it is more robust against potential outliers. The robust correlation coefficient is a weighted correlation coefficient, in which the weights were set to zero for data points judged as potential outliers, based on the residuals of the linear regression (Abdullah, 1990; Eqs. 2.3 and 2.4). The significance was examined using a bootstrap test (10,000 bootstrapped datasets, allowing replacements; one-tailed test).

To ensure the results of the cross-validated ROI analyses (Figure 4B), we also performed additional ROI analyses by orthogonalizing each variable to the variable used to define the ROI, when it was from the same task. Results of these analyses are essentially the same as those shown in Figure 4B. Specifically, an ROI defined by RP in the Control task contained signals significantly modulated by RPE, even when the regressor variable of RPE was orthogonalized to RP in the Control task ($P<0.005$); an ROI defined by RPE contained signals significantly modulated by the regressor variable of RP in the Control task, even when the regressor variable was orthogonalized to RPE ($P<0.005$); an ROI defined by RP in the Other task contained signals significantly modulated by the regressor variable of sRPE, even when the regressor variable was orthogonalized to RP in the Other task ($P<0.005$); and an ROI defined by sRPE contained signals significantly modulated by the regressor variable of RP in the Other task, even when the regressor variable was orthogonalized to sRPE ($P<0.00005$).

Supplemental References

Abdullah, M.B. (1990). On a Robust Correlation Coefficient. J. R. Stat. Soc. Ser. D Appl. Stat. *39*, 455-460.

Behrens, T.E.J., Hunt, L.T., Woolrich, M.W., and Rushworth, M.F.S. (2008). Associative learning of social value. Nature *456*, 245-249.

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

Bishop, C.M. (2006). Pattern Recognition And Machine Learning (Information Science and Statistics) (Springer-Verlag).

Boorman, E.D., Behrens, T.E.J., 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).

Cox, R.W. (1996). AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. Comput. Biomed. Res. *29*, 162-173.

Daw, N.D. (2009). Trial-by-trial data analysis using computational models. Attention and Performance *XXIII*, 26.

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

Efron, B. (1992). Jackknife-after-Bootstrap Standard Errors and Influence Functions. J. R. Stat. Soc. Series. B Stat. Methodol. *54*, 83-127.

Gläscher, J., Daw, N., 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.

Hu, X., Le, T.H., Parrish, T., and Erhard, P. (1995). Retrospective estimation and correction of physiological fluctuation in functional MRI. Magn. Reson. Med. *34*, 201-212.

Huettel, S.A., Stowe, C.J., Gordon, E.M., Warner, B.T., and Platt, M.L. (2006). Neural signatures of economic preferences for risk and ambiguity. Neuron *49*, 765-775.

Kriegeskorte, N., Simmons, W.K., Bellgowan, P.S., and Baker, C.I. (2009). Circular analysis in systems neuroscience: the dangers of double dipping. Nat. Neurosci. *12*, 535-540.

O'Doherty, J.P., Hampton, A., and Kim, H. (2007). Model-Based fMRI and Its Application to Reward Learning and Decision Making. Ann. N. Y. Acad. Sci. *1104*, 35-53.

Stephan, K.E., Penny, W.D., Daunizeau, J., Moran, R.J., and Friston, K.J. (2009). Bayesian model selection for group studies. NeuroImage *46*, 1004-1017.

Sutton, R.S., and Barto, A.G. (1998). Reinforcement Learning: An Introduction (The MIT Press).

Talairach, J., and Tournoux, P. (1988). Co-Planar Stereotaxic Atlas of the Human Brain (Georg Thieme Verlag).

Tobler, P.N., Christopoulos, G.I., O'Doherty, J.P., Dolan, R.J., and Schultz, W. (2009). Risk-dependent reward value signal in human prefrontal cortex. Proc. Natl. Acad. Sci. USA *106*, 7185-7190.