Neuron
Article

Learning to Simulate Others' Decisions

Shinsuke Suzuki,[1] Norihiro Harasawa,[1] Kenichi Ueno,[2] Justin L. Gardner,[3] Noritaka Ichinohe,[5] Masahiko Haruno,[6] Kang Cheng,[2,4] and Hiroyuki Nakahara[1,7,*]
[1]Laboratory for Integrated Theoretical Neuroscience
[2]Support Unit for Functional Magnetic Resonance Imaging
[3]Gardner Research Unit
[4]Laboratory for Cognitive Brain Mapping
RIKEN Brain Science Institute, Wako Saitama, 351-0198, Japan
[5]Department of Ultrastructural Research, National Institute of Neuroscience, NCNP, Kodaira Tokyo, 187-8502, Japan
[6]Center for Information and Neural Networks, NICT, Suita Osaka, 565-0871, Japan
[7]Department of Computational Intelligence & Systems Science, Tokyo Institute of Technology, Yokohama Kanagawa, 226-8503, Japan
*Correspondence: hiro@brain.riken.jp
DOI 10.1016/j.neuron.2012.04.030

SUMMARY

A fundamental challenge in social cognition is how humans learn another person's values to predict their decision-making behavior. This form of learning is often assumed to require simulation of the *other* by direct recruitment of one's own valuation process to model the other's process. However, the cognitive and neural mechanism of simulation learning is not known. Using behavior, modeling, and fMRI, we show that simulation involves two learning signals in a hierarchical arrangement. A simulated-other's reward prediction error processed in ventromedial prefrontal cortex mediated simulation by direct recruitment, being identical for valuation of the self and simulated-other. However, direct recruitment was insufficient for learning, and also required observation of the other's choices to generate a simulated-other's action prediction error encoded in dorsomedial/dorsolateral prefrontal cortex. These findings show that simulation uses a core prefrontal circuit for modeling the other's valuation to generate prediction and an adjunct circuit for tracking behavioral variation to refine prediction.

INTRODUCTION

A fundamental human ability in social environments is the simulation of another person's mental states, or hidden internal variables, to predict their actions and outcomes. Indeed, the ability to simulate another is considered a basic component of mentalizing or theory of mind (Fehr and Camerer, 2007; Frith and Frith, 1999; Gallagher and Frith, 2003; Sanfey, 2007). However, despite its importance for social cognition, little is known about simulation learning and its cognitive and neural mechanisms. A commonly assumed account of simulation is the direct recruitment of one's own decision-making process to model the *other*'s process (Amodio and Frith, 2006; Buckner and Carroll, 2007; Mitchell, 2009). The direct recruitment hypothesis predicts that one makes and simulates a model of how the other will act, including the other's internal variables, as if it is one's own process, and assumes that this simulated internal valuation process employs the same neural circuitry that one uses for one's own process. As such, the hypothesis is parsimonious and thus attractive as a simple explanation of simulation, but it is also difficult to examine experimentally and therefore lies at the heart of current debate in the social cognition literature (Adolphs, 2010; Buckner and Carroll, 2007; Keysers and Gazzola, 2007; Mitchell, 2009; Saxe, 2005). A definitive examination of this issue requires a theoretical framework that provides quantitative predictions that can be tested experimentally.

We adopted a reinforcement learning (RL) framework to provide a simple, rigorous account of behavior in valuating options for one's own decision-making. RL also provides a clear model of one's internal process using two key internal variables: value and reward prediction error. Value is the expected reward associated with available options, and is updated by feedback from a reward prediction error—the difference between the predicted and actual reward. The RL framework is supported by considerable empirical evidence including neural signals in various cortical and subcortical structures that behave as predicted (Glimcher and Rustichini, 2004; Hikosaka et al., 2006; Rangel et al., 2008; Schultz et al., 1997).

The RL framework or other parametric analyses have also been applied to studies of decision making and learning in various social contexts (Behrens et al., 2008; Bhatt et al., 2010; Coricelli and Nagel, 2009; Delgado et al., 2005; Hampton et al., 2008; Montague et al., 2006; Yoshida et al., 2010). These studies investigated how human valuation and choice differ depending on social interactions with others or different understandings of others. They typically require that subjects use high-level mentalizing, or recursive reasoning in interactive game situations where one must predict the other's behavior and/or what they are thinking about themselves. Although important in human social behavior (Camerer et al., 2004; Singer and Lamm, 2009), this form of high-level mentalizing complicates investigation of the signals and computations of simulation and thus makes it difficult to isolate its underlying brain signals.

In the present study, we exploited a basic social situation for our main task, equivalent to a first level (and not higher level) mentalizing process: subjects were required to predict the other's choices while observing their choices and outcomes without interacting with the other. Thus, in our study, the same RL framework that is commonly used to model one's own process provides a model to define signals and computations relevant to the other's process. We also used a control task in which subjects were required to make their own value-based decisions. Combining these tasks allowed us to directly compare brain signals between one's own process and the "simulated-other's" process, in particular, the signals for reward prediction error in one's own valuation (control task) and the simulated-other's valuation (main task).

Moreover, the main task's simple structure makes it relatively straightforward to use the RL framework to identify additional signals and computations beyond those assumed for simulation by direct recruitment. Strongly stated, the direct recruitment hypothesis assumes that the other's process is simulated by the same cognitive and neural process as one's own, and accordingly, in the main task, the simulation learning would be expected to use only knowledge of the other's outcomes, while a weaker version of the hypothesis would assume only the involvement of the cognitive process. Indeed, in many social situations, one may also observe and utilize the other's decisions or choices wherein the stronger hypothesis should be rejected. We therefore examined whether an additional, undefined learning signal based on information about the other's choices might also be used by humans to simulate the other's valuation process.

Employing behavior, fMRI, and computational modeling, we examined the process of simulation learning, asking whether one uses reward prediction errors in the same manner that one does for self learning, and whether the same neural circuitry is recruited. We then investigated whether humans utilize signals acquired by observing variation in the other's choices to improve learning for the simulation and prediction of the other's choice behavior.

RESULTS

Behavior in Simulating the Other's Value-Based Decisions and Making One's Own Decisions

To measure the behavior for learning to simulate the other, subjects performed two decision-making tasks, a Control task and an Other task (Figure 1A). The Other task was designed to probe the subjects' simulation learning to predict the other's value-based decisions, while the Control task was a reference task to probe the subjects' own value-based decisions. In both tasks, subjects repeatedly chose between two stimuli.

In the Control task, only one stimulus was "correct" in each trial, and this was governed by a single reward probability, i.e., the probability p was fixed throughout a block of trials, and the reward probabilities for both stimuli were given by p and $1 - p$, respectively. When subjects made a correct choice, they received a reward with a magnitude that was visibly assigned to the chosen stimulus. As the reward probability was unknown to them, it had to be learned over the course of the trials to maximize overall reward earnings (Behrens et al., 2007). As the reward magnitude for both stimuli was randomly but visibly assigned in each trial, it was neither possible nor necessary to learn to associate specific reward magnitudes with specific stimuli. In fact, because the magnitudes fluctuated across trials, subjects often chose the stimulus with the lower reward probability, even in later trials.

In the Other task, subjects also chose between two stimuli in each trial, but the aim was not to predict which stimulus would give the greatest reward, but to predict the choices made by another person (the other) who was performing the Control task displayed on a monitor (Figure 1A). Subjects were told that the other was a previous participant of the experiment, but their choices were actually generated from an RL model with a risk-neutral setting. Subjects gained a fixed reward in the trial when their predicted choice matched the other's choice; thus, to predict the other's choices, subjects had to learn the reward probability that the other was learning over the trials.

The subjects' choices in the Control task were well fitted by a basic RL model that combined the reward probability and magnitude to compute the value of each stimulus (Equation 1 in Experimental Procedures) and to generate choice probabilities (Figure S1A available online). Given that the reward magnitude was explicitly shown in every trial, the subjects needed to learn only the reward probability. Thus, the RL model was modified such that the reward prediction error is focused on update of the reward probability (Equation 2), not of value per se, as in an earlier study employing this task (Behrens et al., 2007). The RL model correctly predicted the subjects' choices with >90% accuracy (mean ± SEM: 0.9117 ± 0.0098) and provided a better fit to the choice behavior than models using only the reward probability or magnitude to generate choices ($p < 0.01$, paired t test on Akaike's Information Criterion [AIC] value distributions between the two indicated models [Figure 1D]; see Supplemental Experimental Procedures and Table S1 for more details), which is consistent with the earlier study (Behrens et al., 2007).

To compare the subjects' learning of the reward probability in the Control and Other tasks, we plotted the percentage (averaged across all subjects) of times that the stimulus with the higher reward probability was chosen over the course of the trials (Figure 1B, left) and averaged over all trials (Figure 1B, right). During the Control task, subjects learned the reward probability associated with the stimulus and employed a risk-averse strategy. The percentage of times that the stimulus with the higher reward probability was chosen gradually increased during the early trials (Figure 1B, left, blue curve), demonstrating that subjects learned the stimulus reward probability. The average percentage of all trials in which the higher-probability stimulus was chosen (Figure 1B, right, filled blue circle) was significantly higher than the reward probability associated with that stimulus (Figure 1B, right, dashed line; $p < 0.01$, two-tailed t test). This finding suggests that subjects engaged in risk-averse behavior, i.e., choosing the stimulus more often than they should if they were behaving optimally or in a risk-neutral manner. Indeed, in terms of the fit of the RL model (Supplemental Experimental Procedures), the majority of subjects (23/36 subjects) employed risk-averse behavior rather than risk-neutral or risk-prone behavior.

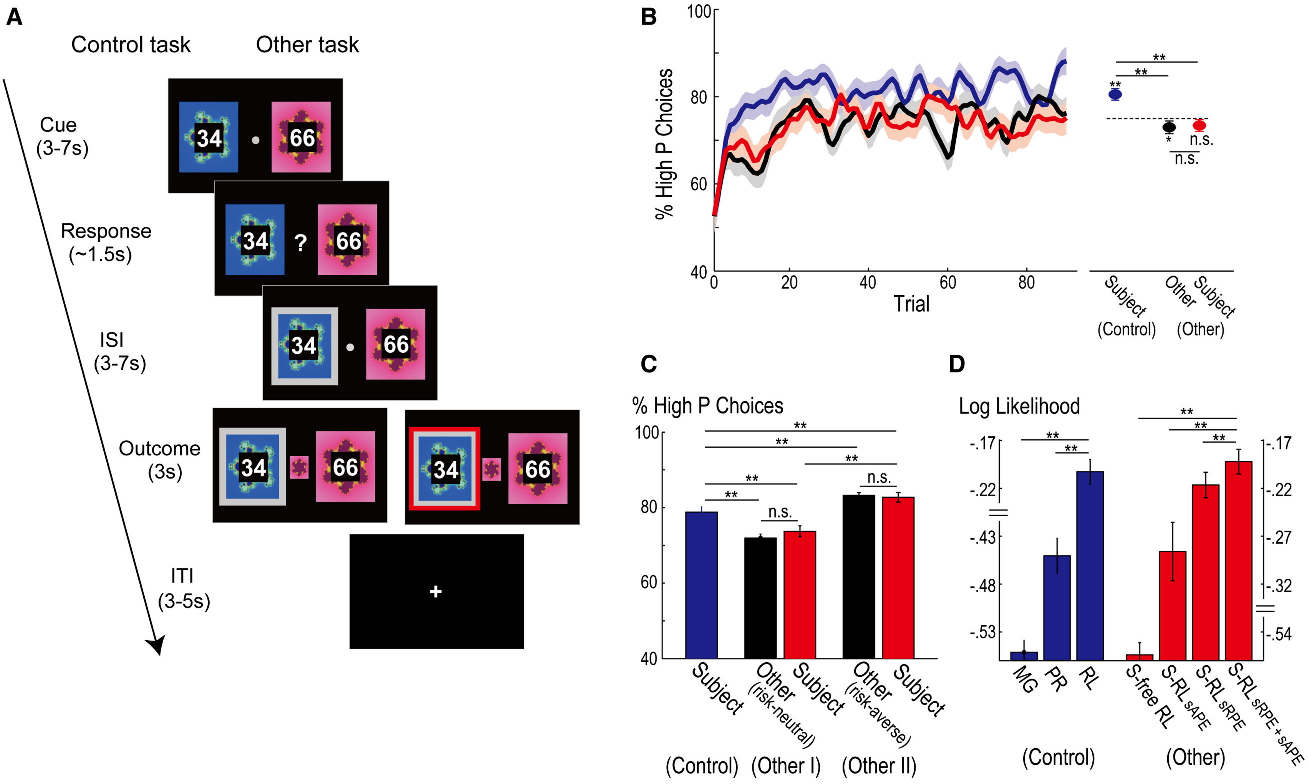

Figure 1. Experimental Tasks and Behavioral Results

(A) Illustration of the experimental tasks: Control (left) and Other (right). In both tasks, each trial consisted of four phases: CUE, RESPONSE, INTERSTIMULUS INTERVAL (ISI), and OUTCOME. For every trial in both tasks, subjects chose between two fractal stimuli, and the stimulus chosen by the subject (RESPONSE) was indicated by a gray frame during the ISI. In the Control task, the "correct" (rewarded) stimulus of the subject was revealed in the center (OUTCOME). In the Other task, the rewarded stimulus of the other was indicated in the center, and the other's choice was indicated by a red frame.

(B) Mean percentages of choosing the stimulus with the higher reward probability (across subjects; n = 36) are shown as curves across trials (left; shaded regions indicate the SEM) and as the averages (±SEM) of all trials (right) for the subjects' choices in the Control (blue) and Other (red) tasks and the others' choices in the Other task (black). These curves were obtained by smoothing each individual's choices with a Gaussian filter (1.25 trials) and then averaging the results for all subjects. The dotted line on the right indicates the stimulus reward probability (75%). Asterisks above the horizontal lines indicate significant differences between the indicated means (**$p < 0.01$; two-tailed paired t test; n.s., nonsignificant as $p > 0.05$), and asterisks at each point indicate significant differences from the stimulus reward probability (*$p < 0.05$, **$p < 0.01$, two-tailed t test; n.s., nonsignificant as $p > 0.05$). Here, we note that the mean percentages of choosing the stimulus with the higher reward probability for the subject and the other in the Other task were slightly lower than the reward probability associated with the stimulus reward probability (subjects: $p = 0.096$; other: $p < 0.05$, two-tailed t test), which is reasonable given that the averaging included the early trials when learning was still ongoing.

(C) Similar data averaged across all trials in a separate experiment (error bars = ± SEM). The two Other task conditions, Other I and Other II, correspond to the other's choices modeled by the RL model using risk-neutral and risk-averse parameters, respectively. **$p < 0.01$, significant differences between the indicated pairs of data (two-tailed paired t test.); n.s., nonsignificant ($p > 0.05$).

(D) Models' fit to behaviors in the Control (left) and Other (right) tasks. 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). **$p < 0.01$, difference in AIC values between the two indicated models (one-tailed paired t test over the AIC distributions). The MG, PR, and RL models in the Control task are the RL model using reward magnitude only, reward probability only, and both, respectively, to generate choices. In the Other task, S-free RL is a simulation-free RL, and S-RL$_{sAPE}$, S-RL$_{sRPE}$, and S-RL$_{sRPE+sAPE}$ are Simulation-RL models using sAPE error only, sRPE only, and both sRPE and sAPE, respectively.

In the Other task, subjects tracked the choice behavior of the other. The percentage of times that the stimulus with the higher reward probability was chosen by the subjects (Figure 1B, left, red curve) appeared to follow the percentage of times that the stimulus was chosen by the other (Figure 1B, left, black curve). This behavior differed from that of the Control task in that the percentage increased over trials but did so more gradually and plateaued at a level below that in the Control task. Indeed, the average percentage of times that the stimulus with the higher reward probability was chosen by the subjects in the Other task (Figure 1B, right, filled red circle) was not significantly different ($p > 0.05$, two-tailed paired t test) from that chosen by the other (Figure 1B, right, filled black circle), but was significantly lower than that chosen by the subjects in the Control task ($p < 0.01$, two-tailed paired t test). Given that the other's choices were modeled using an RL model with a risk-neutral setting, the subjects' choices in the Other task indicate that they were not using risk-averse behavior as they did in the

Control task but were behaving similarly to the other. Together, these results suggest that the subjects were learning to simulate the other's value-based decision making.

Alternative interpretations, however, might also be possible. For example, despite the task instruction to predict the other's choices, the subjects might have completely ignored the other's outcomes and choices and focused instead only on their own outcomes. In this scenario, they might have performed the Other task in the same way as they did the Control task, considering the red frame in the OUTCOME phase (Figure 1A) not as the other's choice, as instructed, but as the ''correct'' stimulus for themselves. Accordingly, such processing can be modeled by reconfiguring the RL model used in the Control task, which is referred to hereafter as simulation-free RL, because it directly associates the options with the outcomes without constructing the other's decision-making process (Dayan and Niv, 2008). This model did not provide a good fit to the behavioral data (see the next section) and can therefore be rejected.

An alternate interpretation is that the subjects focused only on the other's outcomes, processing the other's reward as their own reward, which may have allowed them to learn the reward probability from the assumed reward prediction error. But if this were true, there should have been no difference in their choice behavior between the Control and Other tasks. However, their choice behavior in the Control task was risk-averse and risk-neutral in the Other task, thus refuting this scenario. Nonetheless, it can still be argued that processing the other's reward as their own might have caused the difference in risk behavior between the two tasks; processing the other's reward as their own could have somehow suppressed the risk-averse tendency that existed when they performed for their own rewards, thereby rendering their choice behavior during the Other task similar to the other's risk-neutral behavior. If so, the subjects' choice behavior should always be risk-neutral in the Other task, irrespective of whether or not the other behaves in a risk-neutral manner.

We tested this prediction using another version of the Other task in which the other was modeled by an RL model with a risk-averse setting, and found that, contrary to the prediction, the subjects' behavior tracked that of the Other (Figure 1C). We conducted an additional experiment, adding this ''risk-averse'' Other task as a third task. The subjects' behavior in the original two tasks replicated the findings of the original experiment. Their choices in the third task, however, did not match those made when the other was modeled by the risk-neutral RL model ($p < 0.01$, two-tailed paired t test), but followed the other's choice behavior generated by the risk-averse RL model ($p > 0.05$, two-tailed paired t test). Moreover, the subjects' answers to a postexperiment questionnaire confirmed that they paid attention to both the outcomes and choices of the other (Supplemental Experimental Procedures). These results refute the above argument, and lend support to the notion that the subjects learned to simulate the other's value-based decisions.

Fitting Reinforcement Learning Models for Simulating the Other's Decision-Making Process to Behavior during the Other Task

To determine what information subjects used to simulate the other's behavior, we fitted various computational models simulating the other's value-based decision making to the behavioral data. The general form of these ''simulation-based'' RL models was that subjects learned the simulated-other's reward probability by simulating the other's decision making process. At the time of decision, subjects used the simulated-other's values (the simulated-other's reward probability multiplied by the given reward magnitude) to generate the simulated-other's choice probability, and from this, they could generate their own option value and choice. As discussed earlier, there are two potential sources of information for subjects to learn about the other's decisions, i.e., the other's outcomes and choices.

If subjects applied only their own value-based decision making process to simulate the other's decisions, they would update their simulation using the other's outcomes; they would update the simulated-other's reward probability according to the difference between the other's actual outcome and the simulated-other's reward probability. We termed this difference the ''simulated-other's reward prediction error'' (sRPE; Equation 4).

However, subjects may also use the other's choices to facilitate their learning of the other's process. That is, subjects may also use the discrepancy in their prediction of the other's choices from their actual choices to update their simulation. We termed the difference between the other's choices and the simulated-other's choice probability the ''simulated-other's action prediction error'' (sAPE; Equation 6). In particular, we modeled the sAPE signal as a signal comparable to the sRPE, with the two being combined (i.e., multiplied by the respective learning rates and then added together; Equation 3) to update the simulated-other's reward probability (see Figure S1A for a schematic diagram of the hypothesized computational processes). Computationally, this is achieved such that the sAPE is obtained by transforming the action prediction error that was generated first at the ''action'' level (as the difference between the other's choice and the simulated-other's choice probability [Equation 5; Supplemental Experimental Procedures for more details]) back into the value level.

With these considerations, we examined three simulation-based RL models that learned the simulated-other's reward probability: a model using the sRPE and sAPE (Simulation-$\text{RL}_{\text{sRPE+sAPE}}$), a model using only the sRPE (Simulation-RL_{sRPE}), and a model using only the sAPE (Simulation-RL_{sAPE}). As part of the comparison, we also examined the simulation-free RL model mentioned above.

By fitting each of these computational models separately to the behavioral data and comparing their goodness of fit (Figure 1D; Table S1 for parameter estimates and pseudo-R^2 of each model), we determined that the Simulation-$\text{RL}_{\text{sRPE+sAPE}}$ model provided the best fit to the data. First, all three Simulation-RL models fitted the actual behavior significantly better than the simulation-free RL model ($p < 0.0001$, one-tailed paired t test over the distributions of AIC values across subjects). This broadly supports the notion that subjects took account of and internally simulated the other's decision-making processes in the Other task. Second, the Simulation-$\text{RL}_{\text{sRPE+sAPE}}$ model (S-$\text{RL}_{\text{sRPE+sAPE}}$ model hereafter) fitted the behavior significantly better than the Simulation-RL models using either of the prediction errors alone ($p < 0.01$, one-tailed paired t test over the AIC distributions; Figure 1D). This observation was also supported when examined using other types of statistics: AIC values, a Bayesian comparison using the so-called Bayesian exceedance probability, and the fit of a model of all the subjects together (Table S2). The S-RL$_{sRPE+sAPE}$ model successfully predicted >90% (0.9309 ± 0.0066) of the subjects' choices. Furthermore, as expected from the behavioral results summarized above, only three subjects (3/36) exhibited risk-averse behavior when fit to the S-RL$_{sRPE+sAPE}$ model.

In separate analyses, we confirmed that the sRPE and sAPE provided different information, and that both had an influence on the subjects' predictions of the other's choices. First, both errors (and also their learning rates), as well as the information of the other's actions and choices, were mostly uncorrelated (Supplemental Information), indicating that separate contributions of the two errors are possible. Second, the subjects' choice behavior was found to change in relation to the sAPE (large or small) and the sRPE (positive or negative) in the previous trials and not to the combination of both (two-way repeated-measures ANOVA: $p < 0.001$ for the sRPE main effect, $p < 0.001$ for the sAPE main effect, $p = 0.482$ for their interaction; Figure S1B). This result provides behavioral evidence for separate contributions of the two errors to the subjects' learning.

We next compared the S-RL$_{sRPE+sAPE}$ model to several of its variants. We first examined whether including risk parameters at different levels affected the above finding. The original S-RL$_{sRPE+sAPE}$ model included the risk parameter only in the simulated-other's level (computing the simulated-other's choice probability), but it is possible to consider two other variants of this model: one including a risk parameter only in the subject's level (computing the subject's choice probability) and another including risk parameters in the subject's and simulated-other's levels. Goodness-of-fit comparisons of the original S-RL$_{sRPE+sAPE}$ model with these variants supported the use of the original model (see the Supplemental Information). We then examined the performance of another type of variant, utilized in a recent study (Burke et al., 2010), that used the sAPE not for learning but for biasing the subject's choices in the next trial (Supplemental Experimental Procedures). Comparison of goodness of fit between this variant and the original S-RL$_{sRPE+sAPE}$ model supported the superior fit of the original model ($p < 0.001$, one-tailed paired t test). These results suggest that the subjects learned to simulate the other's value-based decision-making processes using both the sRPE and sAPE.

Neural Signals Reflecting the Simulated-Other's Reward and Action Prediction Errors

We next analyzed fMRI data to investigate which brain regions were involved in simulating the other's decision making processes. Based on the fit of the S-RL$_{sRPE+sAPE}$ model to the behavior in the Other task, we generated regressor variables of interest, including the subject's reward probability at the time of decision (DECISION phase; Materials and Methods) and both the sRPE and sAPE at the time of outcome (OUTCOME phase), and entered them into our whole-brain regression analysis. Similarly, fMRI data from the Control task were analyzed using regressor variables based on the fit of the RL model to the subjects' behavior.

BOLD responses that significantly correlated with the sRPE were found only in the bilateral ventromedial prefrontal cortex (vmPFC; $p < 0.05$, corrected; Figure 2A; Table 1). When these signals were extracted using the leave-one-out cross-validation procedure to provide an independent criterion for region of interest (ROI) selection and thus ensure statistical validity (Kriegeskorte et al., 2009), and then binned according to the sRPE magnitude, the signals increased as the error increased (Spearman's correlation coefficient: 0.178, $p < 0.05$; Figure 2B). As expected for the sRPE, vmPFC signals were found to be positively correlated with the other's outcome and negatively correlated with the simulated-other's reward probability (Figure S2A). As activity in the vmPFC is often broadly correlated with value signals and "self" reward prediction error (Berns et al., 2001; O'Doherty et al., 2007), we further confirmed that the vmPFC signals truly corresponded to the sRPE and were not induced by other variables. The vmPFC signals remained significantly correlated with the sRPE ($p < 0.05$, corrected) even when the following potential confounders were added to our regression analysis: the simulated-other's reward probability, the simulated-other's value for the stimulus chosen by the other as well as by the subject, and the subject's own reward prediction error and reward probability. The vmPFC signals also remained significant even when the regressor variable of the sRPE was first orthogonalized to the sAPE and then included in the regression analysis ($p < 0.05$, corrected). Finally, instead of using the original sRPE, we used the error with the reward magnitude (i.e., the sRPE multiplied by the reward magnitude of the stimulus chosen by the other in each trial) as a regressor in whole-brain analysis. The vmPFC was the only brain area showing activity that was significantly correlated with this error ($p < 0.05$, corrected). These results suggest that activity in the vmPFC exclusively contained information about the sRPE.

The sAPE was significantly correlated with changes in BOLD signals in the right dorsomedial prefrontal cortex (dmPFC; $p < 0.05$, corrected), the right dorsolateral prefrontal cortex (dlPFC; $p < 0.05$, corrected; Figure 2C), and several other regions (Table 1). The dmPFC/dlPFC activity continued to be significantly correlated with the action prediction error, even after cross-validation (dmPFC: 0.200, $p < 0.05$; dlPFC: 0.248, $p < 0.05$; Figure 2D). The dmPFC/dlPFC signals remained significant when potential confounders (the simulated-other's reward probability of the stimulus chosen by the other as well as by the subject) were added to the regression analyses ($p < 0.05$, corrected) or when the regressor variable of the sAPE was first orthogonalized to the sRPE and then included in the regression analysis ($p < 0.05$, corrected). We also confirmed significant activation in the dmPFC/dlPFC ($p < 0.05$, corrected) even when the action prediction error at the action level was used as a regressor variable instead of the error at the value level. The dmPFC/dlPFC areas with significant activation considerably overlapped with the areas originally associated with the significant activation, using the error at the value level (Figure S2B).

Given these findings, we further hypothesized that if the neuronal activity in these brain regions encodes the sRPE and sAPE, then any variability in these signals across subjects should affect their simulation learning and should therefore be reflected in the variation in updating the simulated-other's value using

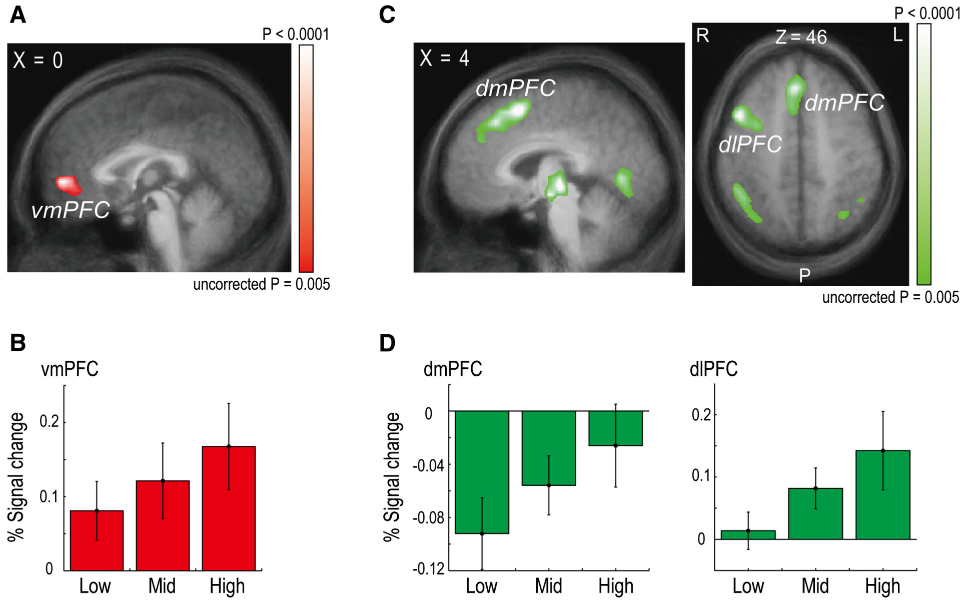

Figure 2. Neural Activity Correlated with the Simulated-Other's Reward and Action Prediction Errors

(A) Neural activity in the vmPFC correlated significantly with the magnitude of the sRPE at the time of outcome (Talairach coordinates: x = 0, y = 53, z = 4). The maps in (A) and (C) are thresholded at $p < 0.005$, uncorrected for display.

(B) Crossvalidated, mean percent changes in the BOLD signals in the vmPFC (across subjects, n = 36; error bars = ± SEM; 7–9 s after the onset of the outcome) during trials in which the sRPE was low, medium, or high (the 33rd, 66th, or 100th percentiles, respectively).

(C) Neural activity in the dmPFC (x = 6, y = 14, z = 52) and dlPFC (x = 45, y = 11, z = 43) correlated significantly with the magnitude of the sAPE at the time of outcome (left: sagittal view; right: axial view).

(D) Crossvalidated, mean percent changes in the BOLD signals in the dmPFC and dlPFC (7–9 s after the onset of the outcome) during trials in which the sAPE was low, medium, or high.

these errors. In other words, subjects with larger or smaller neural signals in a ROI should exhibit larger or smaller behavioral learning effects due to the error (i.e., display larger or smaller learning rates associated with each error).

To test this hypothesis, we investigated the subjects' group-level correlations (Figure 3). Individual differences in the vmPFC BOLD signals of the sRPE (measured by the estimated magnitude of the error's regressor's coefficient; called the "effect size") were correlated with individual differences in the learning rates of the sRPE (determined by the fit of the $\text{S-RL}_{\text{sRPE+sAPE}}$ model to the behavioral data), while those in the dmPFC/dlPFC BOLD signals of the sAPE were correlated with those in the learning rates of the sAPE. First, the vmPFC activity was significantly correlated with the learning rate of the sRPE (Figure 3A, left; Spearman's $\rho = 0.360$, $p < 0.05$), even though the explained variance was relatively small (measured by the square of Pearson's correlation coefficient, $r^2 = 0.124$). We conducted two additional analyses to guard against potential subject outliers that may have compounded the original correlation analysis. The correlation remained significant even when removing all outliers by a Jackknife outlier detection method ($\rho = 0.447$, $p < 0.005$) or using the robust correlation coefficient ($r' = 0.346$, $p < 0.05$) (Supplemental Experimental Procedures). Thus, the observed modulation of vmPFC activity lends correlative support to our hypothesis that variations in the vmPFC signals (putative signals of the sRPE) are associated with the behavioral variability caused by learning using the sRPE across subjects. Second, the dmPFC/dlPFC activity was significantly correlated with the learning rate of the sAPE (Figure 3B, $\rho = 0.330$, $p < 0.05$; $r^2 = 0.140$; and Figure 3C, $\rho = 0.294$, $p < 0.05$; $r^2 = 0.230$). The correlations remained significant after removing the outliers (dmPFC, $\rho = 0.553$, $p < 0.0005$; dlPFC, $\rho = 0.382$, $p < 0.05$) or using the robust correlation coefficient (dmPFC, $r' = 0.377$, $p < 0.005$; dlPFC, $r' = 0.478$, $p < 0.01$). These results support our hypothesis that the variation in the dmPFC and dlPFC signals (putative signals of the sAPE) is associated with the behavioral variability caused by learning using the sAPE across subjects.

Shared Representations of Value-Based Decision Making for the Self and Simulated-Other

We next investigated whether the pattern of vmPFC activity was shared between the self and simulated-other's decision processes in two aspects. First, the vmPFC region was the only region modulated by the sRPE in the Other task. The sRPE was based on simulating the other's process in a social setting, generated in reference to the simulated-other's reward probability that they estimated to substitute for the other's hidden variable. We were then interested in knowing whether the same vmPFC region contained signals for the subject's own reward prediction error during the Control task in a nonsocial setting without the simulation. Second, at the time of decision in the Other task, subjects made their choices to indicate their predictions of the other's choices based on the simulation,

Table 1. Areas Exhibiting Significant Changes in BOLD Signals during the Other Task

Variable	Region	Hemi	BA	x	y	z	t-statistic	p Value
Simulated-other's reward prediction error	**vmPFC**[a]	R/L	10/32	0	53	4	4.45	0.000083
Simulated-other's action prediction error	**dlPFC** (inferior frontal gyrus)	R	44	45	11	43	4.84	0.000026
	dmPFC (medial frontal gyrus/superior frontal gyrus)	R	8	6	14	52	4.73	0.000036
	TPJ/pSTS (inferior parietal lobule/supramarginal gyrus/angular gyrus)	R	39/40	39	−55	37	4.54	0.000064
		L	39/40	−45	−52	37	4.08	0.000246
	Inferior frontal gyrus/superior temporal gyrus	R	47/38	39	20	−5	5.08	0.000013
	Thalamus	R		6	−19	−2	4.88	0.000023
	Lingual gyrus	L	18	12	−73	−8	4.30	0.000131
Reward probability	**vmPFC**	R	10/32	3	56	4	6.16	0.000000
	Postcentral gyrus/superior temporal gyrus	L	2/22/42	−54	−28	16	6.03	0.000001
	Postcentral gyrus/superior temporal gyrus	R	2/22/42	54	−22	19	5.69	0.000002
	Postcentral gyrus	R	1	36	−19	55	5.77	0.000002
	Cingulate gyrus	L	31	−12	−1	34	4.42	0.000092
	Insula	L		−39	−13	4	4.81	0.000028

Activated clusters observed following whole-brain analysis ($p < 0.05$, corrected) of fMRI. The stereotaxic coordinates are in accordance with Talairach space, and the anatomical terms in the Region column are given accordingly. In the far right column, uncorrected p values at the peak of each locus are shown. The regions of interest discussed in the text are shown in bold. vmPFC: ventromedial prefrontal cortex, dlPFC, dorsolateral prefrontal cortex; dmPFC, dorsomedial prefrontal cortex; Hemi, hemisphere; BA, Brodmann area.

[a] The vmPFC region referred to here and in Table 2 is in the vicinity of cluster 2 referred to by Beckmann and colleagues (Beckmann et al., 2009; Rushworth et al., 2011). Upon a closer examination, the locus of the activated vmPFC region is actually located between the BA 10 and 32, and resembles cluster 2, which is also known as area 14 m (Mackey and Petrides, 2010).

whereas in the Control task, they made their choices to obtain the best outcome for themselves without the simulation. Thus, we were also interested in whether the same vmPFC region contained signals for the subjects' decision variables in both types of decisions. To address these issues, we examined the neural correlates of these variables in whole-brain analyses during both tasks and then conducted cross-validating ROI analyses.

We found that the vmPFC was modulated by signals related to the subject's own reward probability in the Other task. Whole-brain analysis during the Other task identified BOLD signals in several brain regions, including the vmPFC ($p < 0.05$, corrected; Figure 4A), that were significantly modulated by the subject's reward probability (for the stimulus chosen by the subject) at the time of decision (Table 1). The subject's reward probability is the decision variable closest to their choices, as it is the farthest downstream in the hypothesized computational processes for generating their choices, but it is also based on simulating the other's decision-making processes, in particular, the simulated-other's reward probability (Figure S1A). To determine whether the activation of the vmPFC that was significantly modulated by the subject's reward probability was compounded by, or possibly rather due to, the simulated-other's reward probability, we conducted two additional whole-brain analyses: when the simulated-other's reward probability (for the stimulus chosen by the subject) was added to the regression analysis as a potential confounder and when the regressor variable of the subject's probability was first orthogonalized to the simulated-other's reward probability and then included in the regression analysis together with the simulated-other's reward probability. In both cases, vmPFC activation remained significantly modulated by the subject's reward probability ($p < 0.05$, corrected). These results indicate that at the time of decision during the Other task, vmPFC activation was significantly modulated by the subject's reward probability.

For comparison, the significant vmPFC signals related to the sRPE are also shown in Figure 4A. Here, we emphasize that the sRPE was not the subject's own reward prediction error (the difference between the subject's own outcome and his/her own reward probability) during the Other task. Indeed, no region was significantly activated by the subject's own reward prediction error during the Other task. This observation was confirmed by an additional whole-brain analysis that was conducted in the same way as the original analysis, except that we added the regressor variable for the subject's own reward prediction error and removed the regressors for the sRPE and sAPE.

Whole-brain analysis during the Control task revealed significant modulation of vmPFC activity ($p < 0.05$, corrected) by the reward probability (for the stimulus chosen by the subject) at the time of the decision and the reward prediction error at the time of the outcome (Figure 4B; Table 2). These activities remained significant ($p < 0.05$, corrected) when the following potential confounders were added to the analysis: the reward magnitude of the chosen stimulus with the reward probability and the value and reward probabilities of the chosen stimulus with the reward prediction error.

We next employed four crossvalidating ROI analyses to investigate whether the same vmPFC region contained signals that were significantly modulated by all four of the variables of interest: the subject's own reward probability (RP) and the sRPE in the Other task (Figure 4A) and the subject's own RP

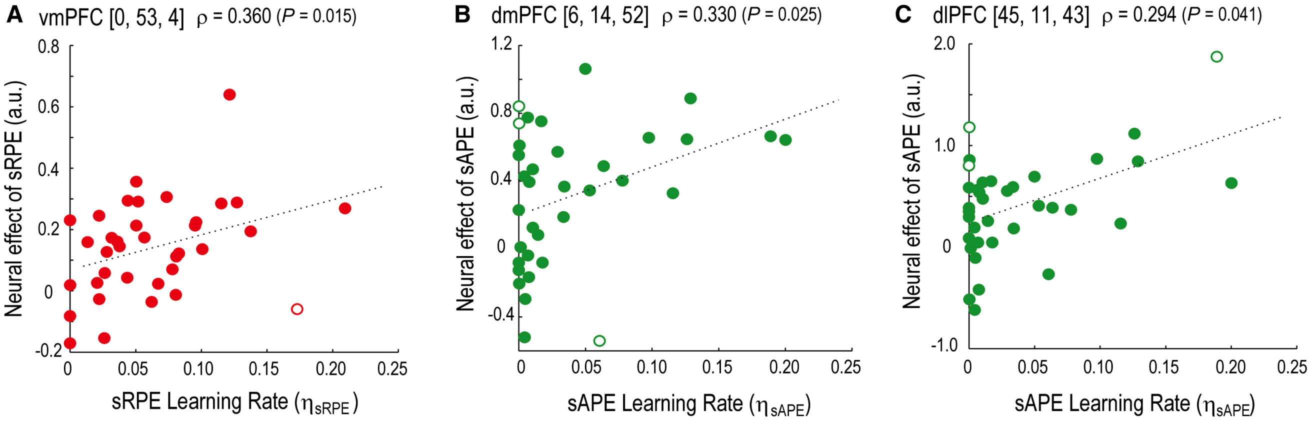

Figure 3. Relationship of Behavioral Variability by Learning Signals with Neural Variability in the vmPFC and the dmPFC/dlPFC

(A) Subject-group level correlation of vmPFC activity for the sRPE with the behavioral effect of the sRPE (the error's learning rate, η_{sRPE}). vmPFC activity is indicated by the error's effect size averaged over the vmPFC region. Open circles denote potential outlier data points (subject) using Jackknife outlier detection.

(B) Correlation of dmPFC activity for the sAPE with the behavioral effect of the sAPE (η_{sAPE}).

(C) Correlation of dlPFC activity for the sAPE with the behavioral effect of the sAPE (η_{sAPE}).

and reward prediction error (RPE) in the Control task (Figure 4B). Whole-brain analyses defined an ROI in the vmPFC for each of these variables. We then examined whether the neural activity in a given ROI was significantly modulated by any or all of the other three variables. Indeed, each of the given ROIs in the vmPFC contained signals that were significantly modulated by each of the variables defining the other three ROIs (either $p < 0.05$ or $p < 0.005$; Figure 4C). We also conducted the same analysis using a Gaussian filter (full width at half-maximum (FWHM) = 6 mm) for spatial smoothing during image data preprocessing that was narrower than the original filter (FWHM = 8 mm). In this case, three of the variables, not RP in the Control task, had significant activation in the vmPFC ($p < 0.05$, corrected; with RP in the Control task, cluster size = 21, which was less than the 33 required for a corrected $p < 0.05$ with the narrower Gaussian filter). However, when the ROI for RP in the Control task was defined under the liberal threshold, we again observed that the activity in a given ROI of one variable was significantly modulated by each of the other three variables ($p < 0.05$). The observation in the original analysis remained true ($p < 0.05$) even if we used an orthogonalized variable in the ROI analysis (see the Supplemental Information). These results indicate that the same region of the vmPFC contains neural signals for the subjects' decisions in both the Control and Other tasks, as well as signals for learning from reward prediction errors either with or without simulation.

DISCUSSION

We examined behavior in a choice paradigm that to our knowledge is new, in which subjects must learn and predict another's value-based decisions. As this paradigm involved observing the other without directly interacting with them, we were able to focus on the most basic form of simulation learning (Amodio and Frith, 2006; Frith and Frith, 1999; Mitchell, 2009). Collectively, our results support the idea of simulation of the other's process by direct recruitment of one's own process, but they also suggest a critical revision to this direct recruitment hypothesis. We found that subjects simultaneously tracked two distinct prediction error signals in simulation learning: the simulated-other's reward and action prediction errors, sRPE and sAPE, respectively. The sRPE significantly modulated signals only in the vmPFC, indicating a prominent role of this area in simulation learning by direct recruitment. However, we also found that simulation learning utilized an accessory learning signal: the sAPE with neural representation in the dmPFC/dlPFC.

Shared Representation between Self and Simulated-Other

Our findings indicate that the vmPFC is a canonical resource for a shared representation between the self and the simulated-other in value-based decision making. By employing a within-subjects design for the Control and Other tasks, the present study provides, to our knowledge, the first direct evidence that vmPFC is the area in which representations of reward prediction error are shared between the self and the simulated-other. Subjects used the sRPE to learn the other's hidden variable and the vmPFC was the only brain region with BOLD signals that were significantly modulated by both the subject's reward prediction error in the Control task and the subject's sRPE in the Other task. Moreover, our findings also provide direct evidence that the same vmPFC region is critical for the subject's decisions, whether or not the other's process was simulated. In both tasks, vmPFC signals were significantly modulated by the subject's decision variable (the subject's reward probability) at the time their decisions were made. Mentalizing by direct recruitment requires the same neural circuitry for shared representations between the self and the simulated-other. Even apart from direct recruitment, shared representations between the self and the other are considered to play an important role in other forms of social cognition, such as empathy. Our findings, with specific roles described for making and learning value-based

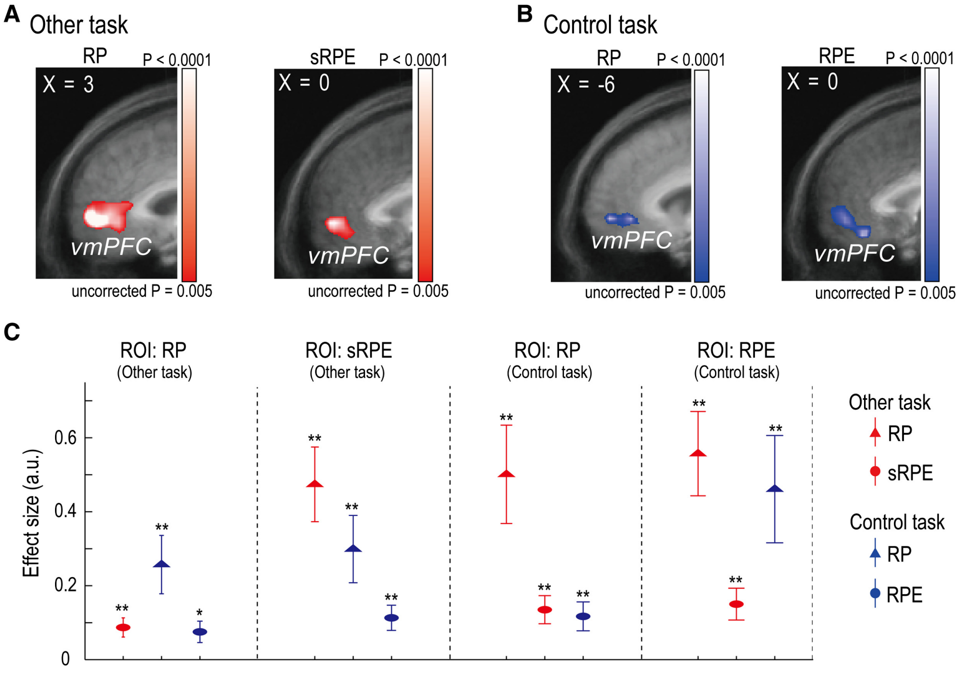

Figure 4. Shared Representations for Self and Other in the vmPFC

(A) (Left) vmPFC signals in the Other task significantly modulated by the subjects' reward probability (RP) at the time of decision (x = 3, y = 56, z = 4; $p < 0.05$, corrected). (Right) The sRPE (x = 0, y = 53, z = 4; $p < 0.05$, corrected) for the signal shown in Figure 2A. The maps in (A) and (B) are thresholded at $p < 0.005$, uncorrected for display.

(B) (Left) vmPFC signals in the Control task significantly modulated by the subjects' reward probability (RP) at the time of DECISION (x = −6, y = 56, z = 1; $p < 0.05$, corrected). (Right) The subjects' reward prediction error at the time of OUTCOME (x = 6, y = 53, z = −2; $p < 0.05$, corrected).

(C) Four ROI analyses showing the extent to which the vmPFC signals represent task-relevant information in the Other (red) and Control (blue) tasks, i.e., RP and sRPE in the Other task and RP and RPE in the Control task. Each plot is labeled with the variable that defined the ROI examined in the vmPFC; the effect sizes of the three other signals on the given ROI are plotted (see symbol legend at right). Points represent the mean (±SEM). *$p < 0.05$, **$p < 0.005$.

decisions, indicate that vmPFC belongs to areas for shared representations in various cognitive domains (Decety and Sommerville, 2003; Keysers and Gazzola, 2007; Mobbs et al., 2009; Rizzolatti and Sinigaglia, 2010; Singer et al., 2004).

For encoding learning signals, the vmPFC is likely more adaptive than the ventral striatum. In contrast to the vmPFC signals, signals in the ventral striatum were significantly modulated only by the subject's own reward prediction error in the Control task (Figure S3; Table 2). The vmPFC was preferentially recruited to simulate the other's process in this study, concordant with the general notion that the vmPFC may encode signals related to reward prediction error when internal models are involved (O'Doherty et al., 2007). The vmPFC may be more sensitive to task demands. During the Other task, no area was significantly modulated by the subject's own reward prediction error. This might be simply due to a limitation in the task design, as the fixed reward size for subjects might have limited detection of reward prediction error. Another aspect, however, is that the subject's own reward prediction error was not as useful as the sRPE for learning to predict the other's choices in this task. Also, the vmPFC may be specifically recruited when subjects used the other's outcomes for learning, as in the Other task, rather than when they vicariously appreciated the other's outcomes. The activity in the ventral striatum might be evoked only when the other's outcomes are more "personal" to subjects (Moll et al., 2006), e.g., when they are comparing their own outcomes to the other's outcomes (Fliessbach et al., 2007; Rilling et al., 2002) or when there are similarities between their

Table 2. Areas Exhibiting Significant Changes in BOLD Signals during the Control Task

Variable	Region	Hemi	BA	x	y	z	*t*-statistic	p Value
Reward prediction error	**vmPFC**	R	10/32	6	53	−2	3.95	0.000360
	ventral striatum	R		(local registration)			4.48	0.000076
Reward probability	**vmPFC**	L	10/32	−6	56	1	4.11	0.000224
	Insula	R		45	−16	7	4.81	0.000028

Activated clusters observed following whole-brain analysis ($p < 0.05$, corrected) of fMRI. Table format is the same as for Table 1. For local registration, see the legend to Figure S3.

own and the other's personal characteristics (Mobbs et al., 2009).

The sRPE was a specific form of reward prediction error related to the other, made in reference to the simulated-other and used for learning their hidden variables. Different forms of the other's reward prediction error also modulated activity in the vmPFC. Activity in the vmPFC was correlated with an "observational" reward prediction error (the difference between the other's stimulus choice outcome and the subject's value of the stimulus) (Burke et al., 2010; Cooper et al., 2011). This error indicated which stimulus was more likely to be rewarding to subjects, whereas in the study presented here, the sRPE indicated which stimulus was more likely to be rewarding to the other. vmPFC signals have also been reported to be modulated by different perceptions of the other's intentions (Cooper et al., 2010). An interesting avenue for future research is to deepen our understanding of the relationship between, and use of, different types of vicarious reward prediction errors involved in forms of fictive or counterfactual learning (Behrens et al., 2008; Boorman et al., 2011; Hayden et al., 2009; Lohrenz et al., 2007).

Refinement of Simulation Learning: Action-Prediction Error

Our findings demonstrate that during simulation, humans use another learning signal—the sAPE—to model the other's internal variables. This error was entirely unexpected based on the direct recruitment hypothesis, and it indicates that simulation is dynamically refined during learning using observations of the other's choices, thus also rejecting the stronger hypothesis.

The sAPE significantly modulated BOLD signals in the dmPFC/dlPFC and several other areas (Table 1), but the sRPE did not. This activation pattern suggests that these areas may have a particular role in utilizing the other's choices rather than the other's outcomes (Amodio and Frith, 2006). This view is convergent with earlier studies in a social context, in which subjects considered the other's behaviors, choices, or intentions, but not necessarily their outcomes (Barraclough et al., 2004; Hampton et al., 2008; Izuma et al., 2008; Mitchell et al., 2006; Yoshida et al., 2010, 2011), and also with studies in nonsocial settings (Gläscher et al., 2010; Li et al., 2011; Rushworth, 2008). Among the other areas, the temporoparietal junction and posterior superior temporal sulcus (TPJ/pSTS) were noteworthy. Our results support a role for the TPJ/pSTS in utilizing the other's choices, consistent with previous studies using RL paradigms in social settings (Behrens et al., 2008; Hampton et al., 2008; Haruno and Kawato, 2009).

Our findings that the dmPFC/dlPFC and TPJ/pSTS were significantly activated by the sAPE in both the value and action levels provide an important twist on the distinction between action and outcome encoding or between action and outcome monitoring (Amodio and Frith, 2006). The signals in those areas represented a result of action monitoring, but were also in a form that was immediately available for learning outcome expectation (the simulated-other's reward probability). It is intriguing to speculate that all of the processes involved in this error, from generating (in the action level) and transforming (from the action to value level) to representing the error as a learning signal for valuation (in the value level), may occur simultaneously in these areas. This would allow the error to be flexibly integrated with other types of processing, thereby leading to better and more efficient learning and decision making (Alexander and Brown, 2011; Hayden et al., 2011).

The sAPE was a specific form of action prediction error related to the other, which was generated in reference to the simulated-other's choice probability and used to learn the simulated-other's variable. Activity in the dmPFC/dlPFC can also be modulated by different forms of action prediction error related to the other and to improvement of the subject's own valuation (Behrens et al., 2008; Burke et al., 2010). Burke et al. (2010) found that activity in the dlPFC was modulated by an observational action prediction error (the difference between the other's actual stimulus choice and the subject's own choice probability). Behrens et al. (2008) found that activity in the dmPFC was significantly modulated by the "confederate prediction error" (the difference between the actual and expected fidelity of the confederate). Their error was used to learn the probability that a confederate was lying in parallel to, but separate from, the learning of the subject's stimulus-reward probability. At the time of decision, subjects could utilize the confederate-lying probability to improve their own decisions. In contrast, in our Other task, subjects needed to predict the other's choices. One possible interpretation is that dmPFC and dlPFC differentially utilize the other's action prediction errors for learning, drawing on different forms of the other's action expectation and/or frames of reference, depending on task demands (Baumgartner et al., 2009; Cooper et al., 2010; de Bruijn et al., 2009; Huettel et al., 2006).

Our findings support a posterior-to-anterior axis interpretation of the dmPFC signals with an increasing order of abstractness to represent the other's internal variable (Amodio and Frith, 2006; Mitchell et al., 2006). The sAPE was in reference to the other's actual choices, whereas the confederate prediction error was in reference to the truth of the other's communicative intentions rather than their choices. Correspondingly, a comparison of the dmPFC regions activated in this study with those in Behrens et al. (2008) suggests that the dmPFC region identified in this study was slightly posterior to the region they identified. Furthermore, our findings also support an axis interpretation between the vmPFC and dmPFC. The sRPE is a more "inner," and thus more abstract, variable for simulation than the sAPE. While the sRPE and sAPE were generated with the simulated-other's reward and choice probability, respectively, this choice probability was generated in each trial by using the reward probability.

Altogether, we propose that the sAPE is a general, critical component for simulation learning. The sAPE provides an additional, but also "natural," learning signal that could arise from simulation by direct recruitment, as it was readily generated from the simulated-other's choice probability given the subject's observation of the other's choices. This error should be useful for refining the learning of the other's hidden variables, particularly if the other behaves differently from the way one would expect for oneself, i.e., the prediction made by direct recruitment simulation (Mitchell et al., 2006). As such, we consider this error and the associated pattern of neural activation to be an accessory signal to the core simulation process of valuation occurring in the vmPFC, which further suggests a more general hierarchy of learning signals in simulation apart from and beyond the sAPE. As the other's choice behavior in this study was only related to a specific personality or psychological isotype, being risk neutral, it will be interesting to see whether and how the sAPE is modified to facilitate learning about the other depending on different personality or psychological isotypes of the other. Also, in this study, because we chose to investigate the sAPE as a general signal, learning about the nature of the other's risk behavior or risk parameters in our model was treated as secondary, being fixed in all trials. However, subjects might have learned the other's risk parameter and/or adjusted their own risk parameter over the course of the trials. How these types of learning complement simulation learning examined in the present study shown here will require further investigation.

Together, we demonstrate that simulation requires distinct prefrontal circuits to learn the other's valuation process by direct recruitment and to refine the overall learning trajectory by tracking the other's behavioral variation. Because our approach used a fundamental form of simulation learning, we expect that our findings may be broadly relevant to modeling and predicting the behavior of others in many domains of cognition, including higher level mentalizing in more complex tasks involving social interactions, recursive reasoning, and/or different task goals. We propose that the signals and computations underlying higher level mentalizing in complex social interactions might be built upon those identified in the present study. It remains to be determined how the simulated-other's reward and action prediction error signals are utilized and modified when task complexity is increased. In this regard, we suggest that the simulation process and the associated neural circuits identified in this study can be conceptualized as a cognitive scaffold upon which multiple context-dependent mentalizing signals may be recruited as available learning signals and may thus contribute to prediction, depending on the subject's goals in the social environment.

EXPERIMENTAL PROCEDURES

We provide a more comprehensive description of the materials and methods in the Supplemental Experimental Procedures.

Subjects

Thirty-nine healthy, normal subjects participated in the fMRI experiment. Subjects received monetary rewards proportional to the points they earned in four test sessions (two fMRI scan sessions, from which behavioral and imaging data are reported in the main text, and two test sessions not involving fMRI, for which data are not shown) in addition to a base participation fee. After excluding three subjects based on their outlier choice behaviors, the remaining 36 subjects were used for subsequent behavioral and fMRI data analyses. A separate behavioral experiment involved 24 normal subjects, and excluding two outlier subjects, the remaining 22 subjects were used for the final analysis (Figure 1C). All subjects gave their informed written consent, and the study was approved by RIKEN's Third Research Ethics Committee.

Experimental Tasks

Two tasks, the Control and Other tasks, were conducted (Figure 1A). The Control task was a one-armed bandit task (Behrens et al., 2007). The two stimuli with randomly assigned reward magnitudes, indicated by numbers in their centers, were randomly positioned at the left or right of the fixation point. In every trial, the reward magnitudes were randomly sampled, independently of the stimuli, but with an additional constraint 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 Figure S1D for control analyses). After subjects made their choice, the chosen stimulus was immediately highlighted by a gray frame. Later, the rewarded stimulus was revealed in the center of the screen. 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 predicted the choice of another person. 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 were generated for the other person performing the Control task. The subjects' prediction of the choice made by the other was immediately highlighted by a gray frame. In the OUTCOME, the other's actual choice was highlighted by a red frame, and the rewarded stimulus for the other was indicated in the center. When the subjects' predicted choice matched the other's actual choice, they earned a fixed reward. The RL model generated the choices of the other on a risk-neutral basis (for the fMRI experiment), so that the choices generated by the model approximately mimicked average (risk-neutral) human behavior, allowing us to use the same type of the other's behavior for all subjects (see Figure S1C for a separate behavioral analysis of this approach).

For the experiment in the MRI scanner, two tasks, Control and Other, were employed. Three conditions, one Control and two Others, were used in a separate behavioral experiment (Figure 1C). The settings for the Control and "Other I" task were the same as in the fMRI experiment, but in the "Other II" task, a risk-averse RL model was used to generate the other's choices.

Behavioral Analysis and Computational Models Fitted to Behavior

Several computational models, based on and modified from the Q learning model (Sutton and Barto, 1998), were fit to the subjects' choice behaviors in both tasks. In the Control task, the RL model, being risk neutral, constructed *Q* values of both stimuli; the value of a stimulus was the product of the stimulus' reward probability, $p(A)$ (for stimulus *A*; the following description is made for this case), and the reward magnitude of the stimulus in a given trial, $R(A)$,

$$Q_A = p(A)R(A). \tag{1}$$

To account for possible risk behavior of the subjects, we followed the approach of Behrens et al. (2007) by using a simple nonlinear function (see the Supplemental Information for more details and for a control analysis of the nonlinear function). The choice probability is given by $q(A) = f(Q_A - Q_B)$, where *f* is a sigmoidal function. The reward prediction error was used to update the stimulus' reward probability (see the Supplemental Information for a control analysis),

$$\delta = r - p(A), \tag{2}$$

where *r* is the reward outcome (1 if stimulus *A* is rewarded and 0 otherwise). The reward probability was updated using $p(A) \leftarrow p(A) + \eta\delta$.

In the Other task, the S-RL$_{sRPE+sAPE}$ model computed the subject's choice probability using $q(A) = f(Q_A - Q_B)$; here, the value of a stimulus is the product of the subject's fixed reward outcome and their reward probability based on simulating the other's decision making, which is equivalent to the simulated-other's choice probability: $q_o(A) = f(Q_O(A) - Q_O(B))$, wherein the other's value of a stimulus is the product of the other's reward magnitude of the stimulus and the simulated-other's reward probability, $p_O(A)$. When the outcome for the other (r_O) was revealed, the S-RL$_{sRPE+sAPE}$ model updated the simulated-other's reward probability, using both the sRPE and the sAPE,

$$p_O(A) \leftarrow p_O(A) + \eta_{sRPE}\delta_O(A) + \eta_{sAPE}\sigma_O(A), \tag{3}$$

where the two η's indicate the respective learning rates. The sRPE was given by

$$\delta_o(A) = r_o - p_o(A). \tag{4}$$

The sAPE was defined in the value level, being comparable to the sRPE. After being generated first in the action level,

$$\sigma'_O(A) = I_A(A) - q_O(A) = 1 - q_O(A), \tag{5}$$

the sAPE was obtained by a variational transformation, pulled back to the value level,

$$\sigma_O(A) = \sigma'_O \frac{(A)}{K}, \qquad (6)$$

(see the Supplemental Information for the algebraic expression of *K*). The two other simulation-RL models only used one of the two prediction errors. The simulation-free RL model is described in the Supplemental Information.

We used a maximum-likelihood approach to fit the models to the individual subject's behaviors and AIC to compare their goodness of fit, taking into account the different numbers of the models' parameters. 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.

fMRI Acquisition and Analysis

fMRI images were collected using a 4 T MRI system (Agilient Inc., Santa Clara, CA). BOLD signals were measured using a two-shot EPI sequence. High- and low-resolution whole-brain anatomical images were acquired using a T1-weighted 3D FLASH pulse sequence. All images were analyzed using Brain Voyager QX 2.1 (Brain Innovation B.V., Maastricht, The Netherlands). Functional images were preprocessed, including spatial smoothing with a Gaussian filter (FWHM = 8 mm). Anatomical images were transformed into the standard Talairach space (TAL) and functional images were registered to high-resolution anatomical images. All activations were reported based on the TAL, except for the activation in the ventral striatum reported in Figure S3 (see legend).

We employed model-based analysis to analyze the BOLD signals. The main variables of interest as the regressors for our regression analyses were, for the Control task, the reward probability of the stimulus chosen in the DECISION period (defined as the period from the onset of CUE until subjects made their responses in the RESPONSE period) and the reward prediction error in the OUTCOME period. For the Other task, the main variables of interest were the subject's reward probability for the stimulus chosen in the DECISION period, and the sRPE and sAPE in the OUTCOME period. Random-effects analysis was employed using a one-tailed t test. Significant BOLD signals were reported based on corrected p values ($p < 0.05$) using a family-wise error for multiple comparison corrections (cluster-level inference). For cross-validated percent changes in the BOLD signals (Figures 2B and 2D), we followed a previously described leave-one-out procedure (Gläscher et al., 2010). For the correlation analysis (Figure 3), we calculated Spearman's correlation coefficient and tested its statistical significance using a one-tailed t test given our hypothesis of positive correlation (see the Supplemental Information for two additional analyses).

SUPPLEMENTAL INFORMATION

Supplemental Information includes three figures, two tables, and Supplemental Experimental Procedures and can be found with this article online at doi:10.1016/j.neuron.2012.04.030.

ACKNOWLEDGMENTS

This work was supported by KAKENHI grants 21300129 and 20020034 (H.N.). We thank S. Kaveri for discussion in the early stages of this work, Dr. X.H. Wan for assistance with data analysis, Drs. K. Tanaka and N. Sadato for helpful comments on the manuscript, and Drs. T. Asamizuya and C. Suzuki for technical assistance with the fMRI experiments.

Accepted: April 10, 2012
Published: June 21, 2012

REFERENCES

Adolphs, R. (2010). Conceptual challenges and directions for social neuroscience. Neuron *65*, 752–767.

Alexander, W.H., and Brown, J.W. (2011). Medial prefrontal cortex as an action-outcome predictor. Nat. Neurosci. *14*, 1338–1344.

Amodio, D.M., and Frith, C.D. (2006). Meeting of minds: the medial frontal cortex and social cognition. Nat. Rev. Neurosci. *7*, 268–277.

Barraclough, D.J., Conroy, M.L., and Lee, D. (2004). Prefrontal cortex and decision making in a mixed-strategy game. Nat. Neurosci. *7*, 404–410.

Baumgartner, T., Fischbacher, U., Feierabend, A., Lutz, K., and Fehr, E. (2009). The neural circuitry of a broken promise. Neuron *64*, 756–770.

Beckmann, M., Johansen-Berg, H., and Rushworth, M.F. (2009). Connectivity-based parcellation of human cingulate cortex and its relation to functional specialization. J. Neurosci. *29*, 1175–1190.

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.

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.

Berns, G., McClure, S., Pagnoni, G., and Montague, P. (2001). Predictability modulates human brain response to reward. J. Neurosci. *21*, 2793–2798.

Bhatt, M.A., Lohrenz, T., Camerer, C.F., and Montague, P.R. (2010). Neural signatures of strategic types in a two-person bargaining game. Proc. Natl. Acad. Sci. USA *107*, 19720–19725.

Boorman, E.D., Behrens, T.E., and Rushworth, M.F. (2011). Counterfactual choice and learning in a neural network centered on human lateral frontopolar cortex. PLoS Biol. *9*, e1001093.

Buckner, R.L., and Carroll, D.C. (2007). Self-projection and the brain. Trends Cogn. Sci. (Regul. Ed.) *11*, 49–57.

Burke, C.J., Tobler, P.N., Baddeley, M., and Schultz, W. (2010). Neural mechanisms of observational learning. Proc. Natl. Acad. Sci. USA *107*, 14431–14436.

Camerer, C.F., Ho, T., and Chong, J. (2004). A cognitive hierarchy model of games*. Q. J. Econ. *119*, 861–898.

Cooper, J.C., Kreps, T.A., Wiebe, T., Pirkl, T., and Knutson, B. (2010). When giving is good: ventromedial prefrontal cortex activation for others' intentions. Neuron *67*, 511–521.

Cooper, J.C., Dunne, S., Furey, T., and O'Doherty, J.P. (2011). Human dorsal striatum encodes prediction errors during observational learning of instrumental actions. J. Cogn. Neurosci. *24*, 106–118.

Coricelli, G., and Nagel, R. (2009). Neural correlates of depth of strategic reasoning in medial prefrontal cortex. Proc. Natl. Acad. Sci. USA *106*, 9163–9168.

Dayan, P., and Niv, Y. (2008). Reinforcement learning: the good, the bad and the ugly. Curr. Opin. Neurobiol. *18*, 185–196.

de Bruijn, E.R.A., de Lange, F.P., von Cramon, D.Y., and Ullsperger, M. (2009). When errors are rewarding. J. Neurosci. *29*, 12183–12186.

Decety, J., and Sommerville, J.A. (2003). Shared representations between self and other: a social cognitive neuroscience view. Trends Cogn. Sci. (Regul. Ed.) *7*, 527–533.

Delgado, M.R., Frank, R.H., and Phelps, E.A. (2005). Perceptions of moral character modulate the neural systems of reward during the trust game. Nat. Neurosci. *8*, 1611–1618.

Fehr, E., and Camerer, C.F. (2007). Social neuroeconomics: the neural circuitry of social preferences. Trends Cogn. Sci. (Regul. Ed.) *11*, 419–427.

Fliessbach, K., Weber, B., Trautner, P., Dohmen, T., Sunde, U., Elger, C.E., and Falk, A. (2007). Social comparison affects reward-related brain activity in the human ventral striatum. Science *318*, 1305–1308.

Frith, C.D., and Frith, U. (1999). Interacting minds—a biological basis. Science *286*, 1692–1695.

Gallagher, H.L., and Frith, C.D. (2003). Functional imaging of 'theory of mind'. Trends Cogn. Sci. (Regul. Ed.) *7*, 77–83.

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.

Glimcher, P.W., and Rustichini, A. (2004). Neuroeconomics: the consilience of brain and decision. Science *306*, 447–452.

Hampton, A.N., Bossaerts, P., and O'Doherty, J.P. (2008). Neural correlates of mentalizing-related computations during strategic interactions in humans. Proc. Natl. Acad. Sci. USA *105*, 6741–6746.

Haruno, M., and Kawato, M. (2009). Activity in the superior temporal sulcus highlights learning competence in an interaction game. J. Neurosci. *29*, 4542–4547.

Hayden, B.Y., Pearson, J.M., and Platt, M.L. (2009). Fictive reward signals in the anterior cingulate cortex. Science *324*, 948–950.

Hayden, B.Y., Heilbronner, S.R., Pearson, J.M., and Platt, M.L. (2011). Surprise signals in anterior cingulate cortex: neuronal encoding of unsigned reward prediction errors driving adjustment in behavior. J. Neurosci. *31*, 4178–4187.

Hikosaka, O., Nakamura, K., and Nakahara, H. (2006). Basal ganglia orient eyes to reward. J. Neurophysiol. *95*, 567–584.

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.

Izuma, K., Saito, D.N., and Sadato, N. (2008). Processing of social and monetary rewards in the human striatum. Neuron *58*, 284–294.

Keysers, C., and Gazzola, V. (2007). Integrating simulation and theory of mind: from self to social cognition. Trends Cogn. Sci. (Regul. Ed.) *11*, 194–196.

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.

Li, J., Delgado, M.R., and Phelps, E.A. (2011). How instructed knowledge modulates the neural systems of reward learning. Proc. Natl. Acad. Sci. USA *108*, 55–60.

Lohrenz, T., McCabe, K., Camerer, C.F., and Montague, P.R. (2007). Neural signature of fictive learning signals in a sequential investment task. Proc. Natl. Acad. Sci. USA *104*, 9493–9498.

Mackey, S., and Petrides, M. (2010). Quantitative demonstration of comparable architectonic areas within the ventromedial and lateral orbital frontal cortex in the human and the macaque monkey brains. Eur. J. Neurosci. *32*, 1940–1950.

Mitchell, J.P. (2009). Inferences about mental states. Philos. Trans. R. Soc. Lond. B Biol. Sci. *364*, 1309–1316.

Mitchell, J.P., Macrae, C.N., and Banaji, M.R. (2006). Dissociable medial prefrontal contributions to judgments of similar and dissimilar others. Neuron *50*, 655–663.

Mobbs, D., Yu, R., Meyer, M., Passamonti, L., Seymour, B., Calder, A.J., Schweizer, S., Frith, C.D., and Dalgleish, T. (2009). A key role for similarity in vicarious reward. Science *324*, 900.

Moll, J., Krueger, F., Zahn, R., Pardini, M., de Oliveira-Souza, R., and Grafman, J. (2006). Human fronto-mesolimbic networks guide decisions about charitable donation. Proc. Natl. Acad. Sci. USA *103*, 15623–15628.

Montague, P.R., King-Casas, B., and Cohen, J.D. (2006). Imaging valuation models in human choice. Annu. Rev. Neurosci. *29*, 417–448.

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.

Rangel, A., Camerer, C., and Montague, P.R. (2008). A framework for studying the neurobiology of value-based decision making. Nat. Rev. Neurosci. *9*, 545–556.

Rilling, J., Gutman, D., Zeh, T., Pagnoni, G., Berns, G., and Kilts, C. (2002). A neural basis for social cooperation. Neuron *35*, 395–405.

Rizzolatti, G., and Sinigaglia, C. (2010). The functional role of the parieto-frontal mirror circuit: interpretations and misinterpretations. Nat. Rev. Neurosci. *11*, 264–274.

Rushworth, M.F. (2008). Intention, choice, and the medial frontal cortex. Ann. N Y Acad. Sci. *1124*, 181–207.

Rushworth, M.F., 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.

Sanfey, A.G. (2007). Social decision-making: insights from game theory and neuroscience. Science *318*, 598–602.

Saxe, R. (2005). Against simulation: the argument from error. Trends Cogn. Sci. (Regul. Ed.) *9*, 174–179.

Schultz, W., Dayan, P., and Montague, P.R. (1997). A neural substrate of prediction and reward. Science *275*, 1593–1599.

Singer, T., and Lamm, C. (2009). The social neuroscience of empathy. Ann. N Y Acad. Sci. *1156*, 81–96.

Singer, T., Seymour, B., O'Doherty, J., Kaube, H., Dolan, R.J., and Frith, C.D. (2004). Empathy for pain involves the affective but not sensory components of pain. Science *303*, 1157–1162.

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

Yoshida, W., Seymour, B., Friston, K.J., and Dolan, R.J. (2010). Neural mechanisms of belief inference during cooperative games. J. Neurosci. *30*, 10744–10751.

Yoshida, K., Saito, N., Iriki, A., and Isoda, M. (2011). Representation of others' action by neurons in monkey medial frontal cortex. Curr. Biol. *21*, 249–253.

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

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

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

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

Table S1 – related to Figure 1: Best fitting parameter estimates

Table S2 – related to Figure 1: Model comparison among Simulation(S)-RL$_{\mathrm{sRPE+sAPE}}$, S-RL$_{\mathrm{sRPE}}$ and S-RL$_{\mathrm{sAPE}}$ models

Supplemental Experimental Procedures

Supplemental Figures

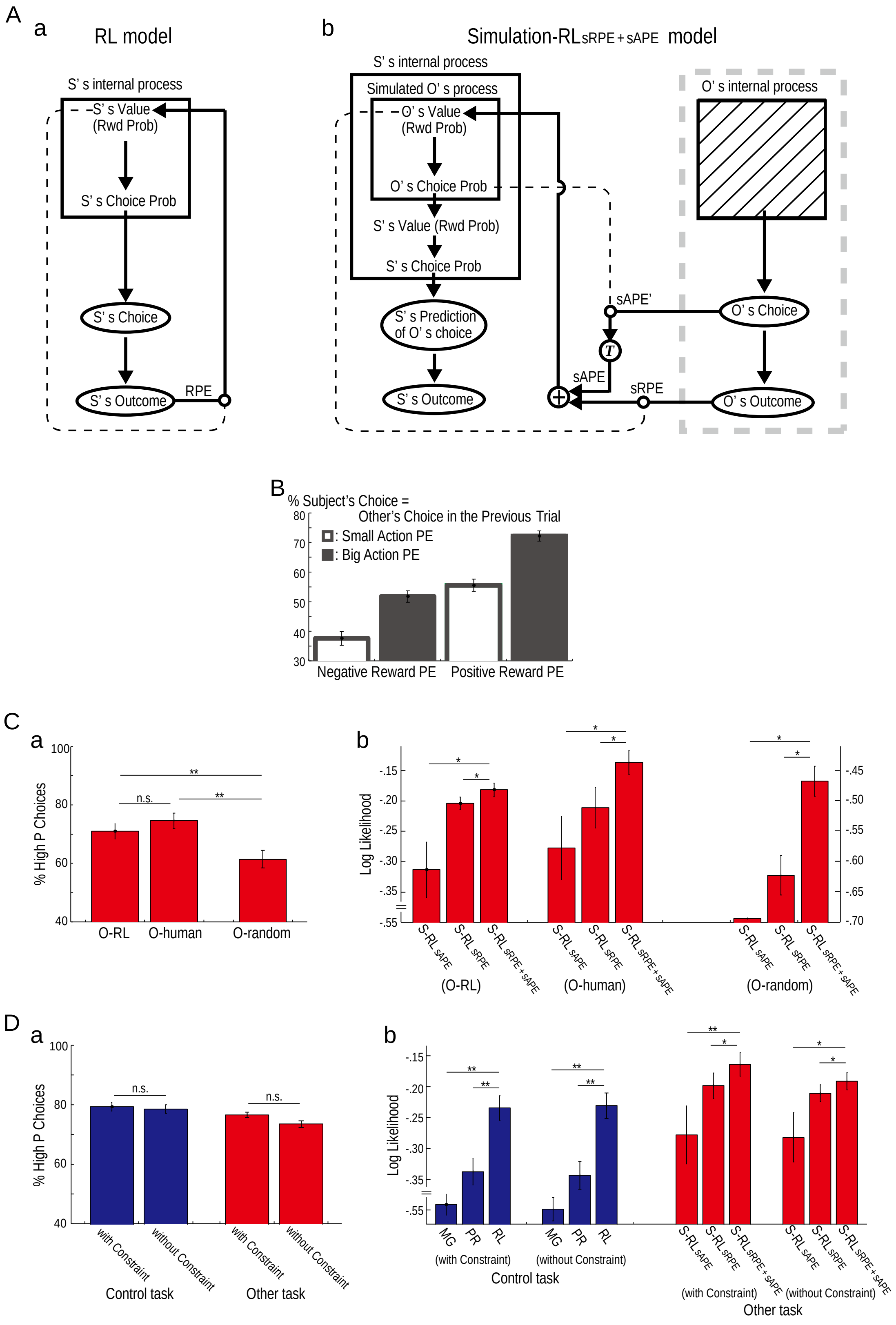

Figure S1 – related to Figure 1: Schematic diagrams of decision making processes used in this study and additional behavioral results

(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$_{\mathrm{sRPE+sAPE}}$ (S-RL$_{\mathrm{sRPE+sAPE}}$) 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$_{\mathrm{sRPE+sAPE}}$, 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$_{\mathrm{sRPE+sAPE}}$ 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.

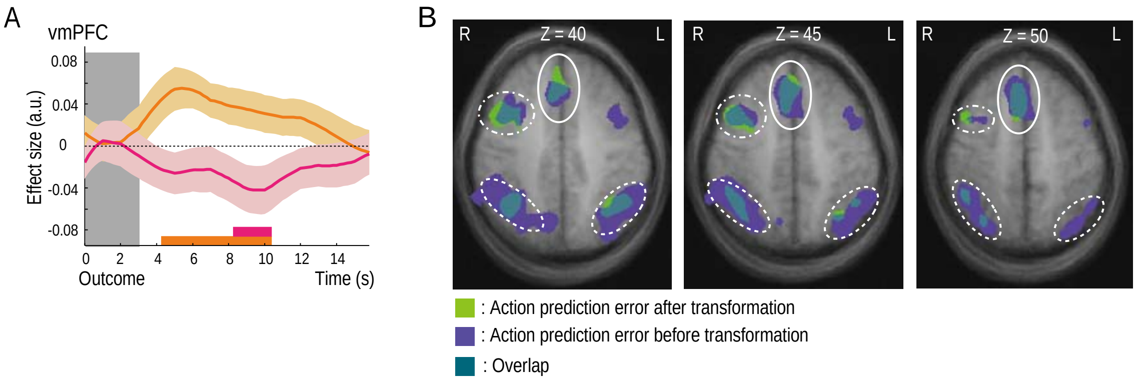

Figure S2 related to Figure 2: Additional results of the neural correlates of simulated other's reward and action prediction errors

(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.

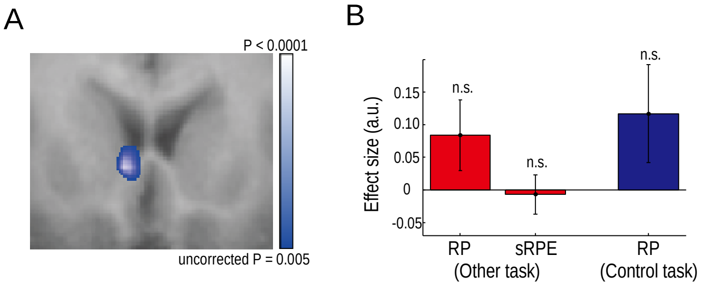

Figure S3 related to Figure 4: Reward prediction error signals in the ventral striatum (vStr) during the Control task.

(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		Stochasticity in the choices, β	Risk parameter γ	pseudo-R^2
		sRPE	sAPE			
$\text{S-RL}_{\text{sRPE+sAPE}}$						
	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
$\text{S-RL}_{\text{sRPE}}$						
	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
$\text{S-RL}_{\text{sAPE}}$						
	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+ $2k\left(k+1\right)/\left(n-k-1\right)$, 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 $w=\eta_{RPE}/\left(\eta_{RPE}+\eta_{APE}\right)$, $[-7,15]$ for $\log\left(\gamma\right)$). 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

$$Q_S(A) = R_S(A) \cdot p_S(A), \tag{1}$$

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

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

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

$$q_S(A) = f\left(Q_S(A) - Q_S(B)\right), \tag{3}$$

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

$$\delta_S = r_S - p_S(A), \tag{4}$$

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$_{\text{sRPE+sAPE}}$ model (S-RL$_{\text{sRPE+sAPE}}$) (see Supplemental Figure S1A). In each trial, the S-RL$_{\text{sRPE+sAPE}}$ 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 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

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

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$_{\text{sRPE+sAPE}}$ 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,

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

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$_{\text{sRPE+sAPE}}$ updated the simulated-other's reward probability, using both the simulated-other's reward and action prediction errors together, given by

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

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$_{\text{sRPE+sAPE}}$ 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 S-RL$_{\text{sRPE+sAPE}}$ differs from the S-RL$_{\text{sRPE}}$ in that it uses the action prediction error as an additional learning signal. Therefore, the 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 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$_{\mathrm{sRPE+sAPE}}$ model to the behavior in the Other task, compared with that of the original S-RL$_{\mathrm{sRPE+sAPE}}$ 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$_{\mathrm{sRPE+sAPE}}$ 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$_{\mathrm{sRPE+sAPE}}$ 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$_{\mathrm{sRPE+sAPE}}$ 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$_{\mathrm{sRPE+sAPE}}$ 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 σ'_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$_{\mathrm{sRPE+sAPE}}$ 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$_{\mathrm{sRPE+sAPE}}$ 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$_{\mathrm{sRPE+sAPE}}$ model; the '1st-R-variant' and '1st-A-variant' S-RL$_{\mathrm{sRPE+sAPE}}$ 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$_{\mathrm{sRPE+sAPE}}$ 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$_{\mathrm{sRPE+sAPE}}$ model, there were three-variants; the '2nd-R-variant' and '2nd-A-variant' of the S-RL$_{\mathrm{sRPE+sAPE}}$ 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$_{\mathrm{sRPE+sAPE}}$ 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$_{\mathrm{sRPE+sAPE}}$ 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$_{\mathrm{sRPE+sAPE}}$ 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$_{\mathrm{sRPE+sAPE}}$, $P < 0.01$).

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.