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.

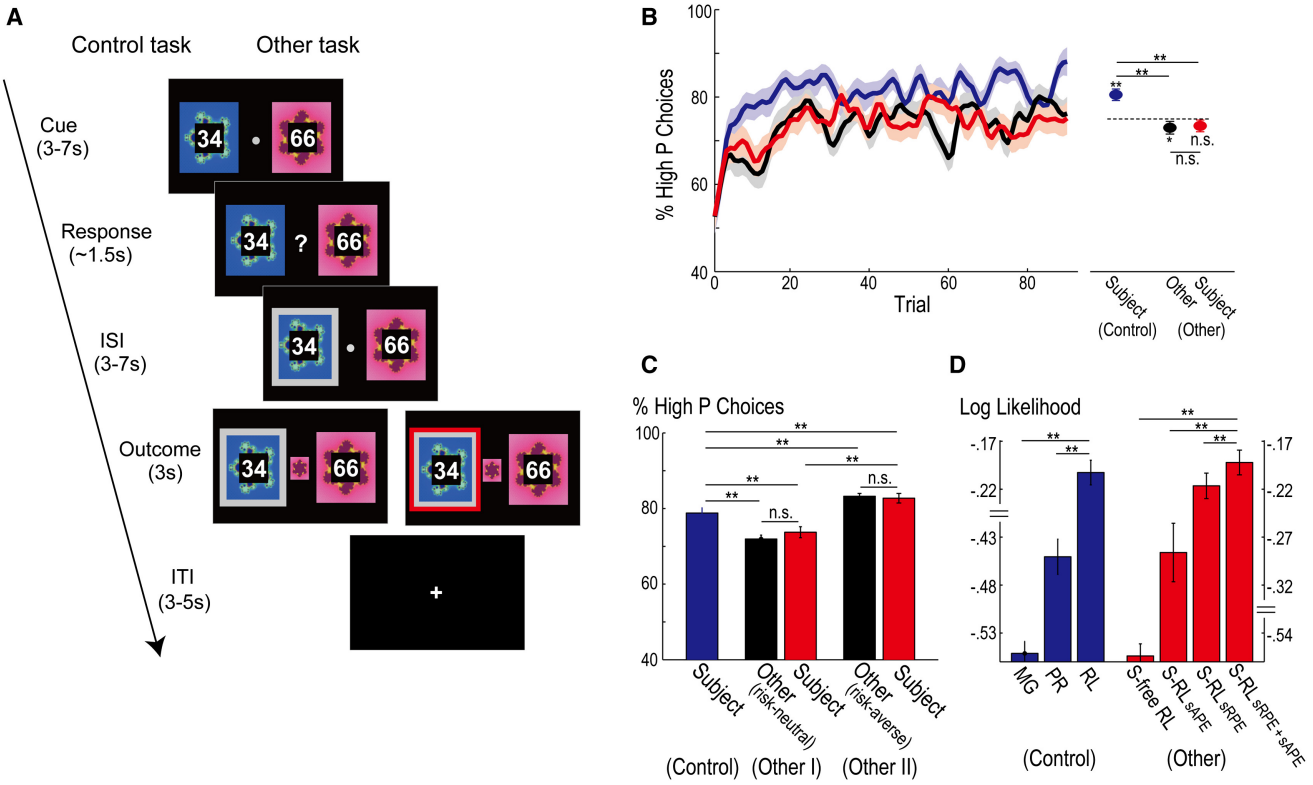

Figure 1. Experimental Tasks and Behavioral Results

(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-$RL_{sRPE+sAPE}$), 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-$RL_{sRPE+sAPE}$ 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-$RL_{sRPE+sAPE}$ model (S-$RL_{sRPE+sAPE}$ 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

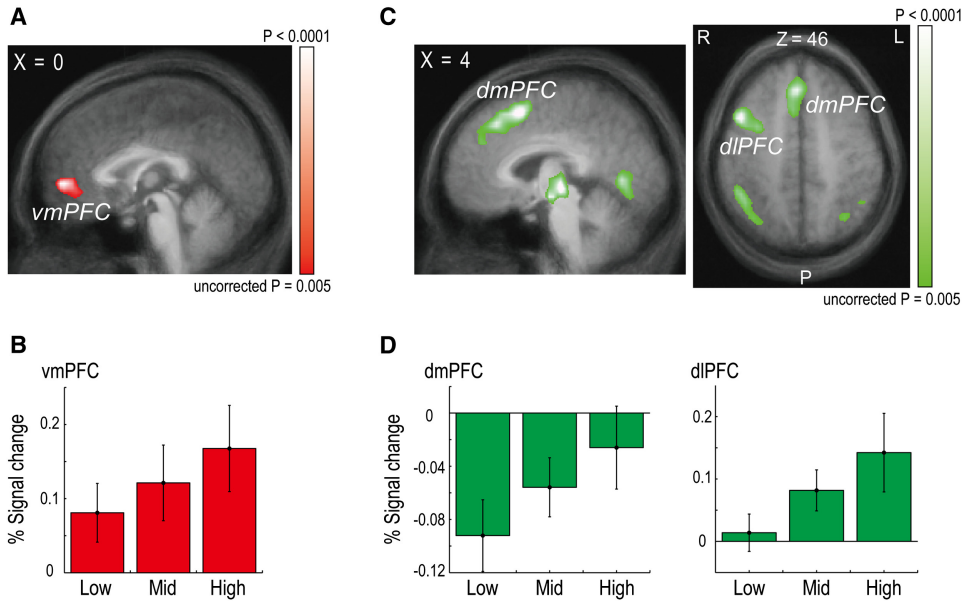

Figure 2. Neural Activity Correlated with the Simulated-Other's Reward and Action Prediction Errors

(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 S-RL$_{sRPE+sAPE}$ 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

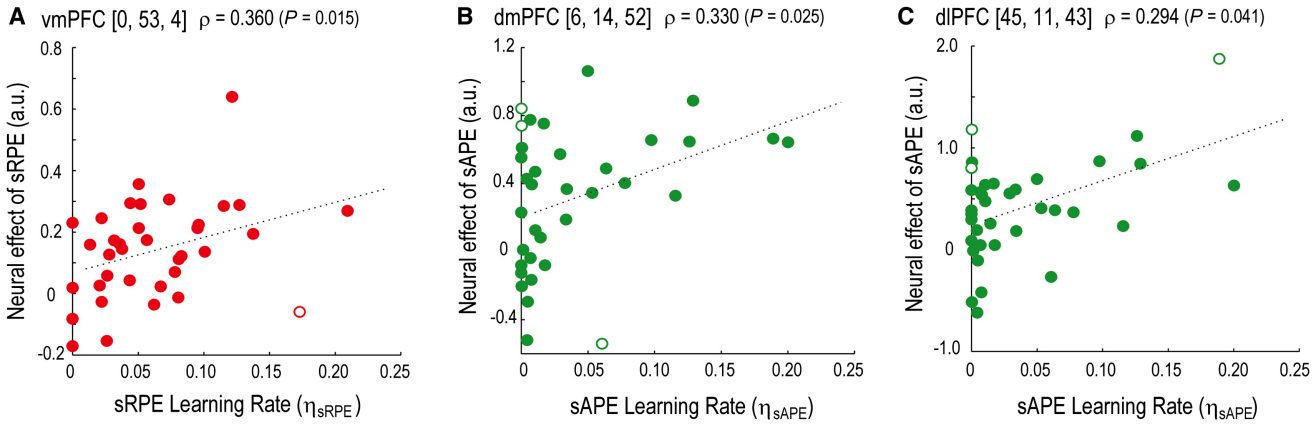

Figure 3. Relationship of Behavioral Variability by Learning Signals with Neural Variability in the vmPFC and the dmPFC/dlPFC

(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

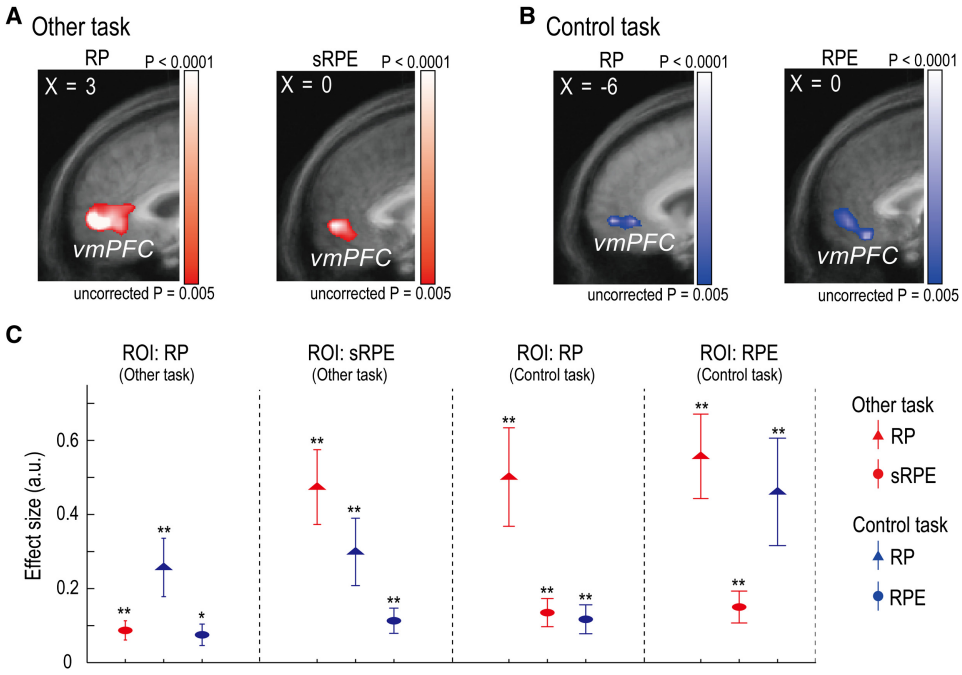

Figure 4. Shared Representations for Self and Other in the vmPFC

(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, $R(A)$,

$$Q_A = p(A)R(A). \quad (1)$$

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),

$$\delta = r - p(A), \quad (2)$$

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$_{\mathrm{sRPE+sAPE}}$ 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$_{\mathrm{sRPE+sAPE}}$ model updated the simulated-other's reward probability, using both the sRPE and the sAPE,

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

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

$$\delta_o(A) = r_o - p_o(A). \quad (4)$$

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

$$\sigma'_O(A) = I_A(A) - q_O(A) = 1 - q_O(A), \quad (5)$$

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

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

(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.