Supporting Information

Zhu et al. 10.1073/pnas.1116783109

SI Methods

Procedure. Before entering the scanner, subjects were given instructions and completed a quiz to ensure comprehension of the game. In the Patent Race, players were matched at random at the beginning of each round and competed for a prize by choosing an investment from their respective endowments. The player who invested more won the prize, and the other lost. In the event of a tie, both lost the prize. Regardless of the outcome, players lost the amount that they invested (Fig. 1). In the particular payoff structure we used, the prize was worth 10 units, and the Strong (Weak) player was endowed with 5 (4) units.

To overcome logistic difficulties of conducting simultaneous experiments with upwards of 16 subjects for each neuroimaging subject, and to minimize unobserved session effects in opponent play associated with such a protocol, we matched subjects with choices from a pool of players who previously participated in behavioral sessions. Importantly, subjects were informed that they played in the same sequence as the pool players. That is, if the scanner subject was playing in round 60, the choice of opponent was drawn randomly from round 60 of one of the pool players (*SI Results*).

fMRI Scanning Parameters. Functional MR images were obtained for each subject by using a 3.0 Tesla Siemens Allegra scanner located at the research-dedicated Beckman Imaging Center (BIC) at the University of Illinois at Urbana–Champaign. Images were acquired by using echo-planar T2* images with BOLD (blood oxygenation-level-dependent) contrast, and angled 30° with respect to the AC-PC line to minimize susceptibility artifacts in the orbitofrontal cortex (1). MR imaging settings were as follows: repetition time (TR) = 2,000 ms; echo time (TE) = 40 ms; slice thickness = 3 mm yielding a 64 × 64 × 32 matrix (3 mm × 3 mm × 3 mm); flip angle = 90°; FOV read = 220 mm; FOV phase = 100 mm, interleaved series order. High-resolution structural T1-weighted scans (1 mm × 1 mm × 1 mm) were acquired by using an MPRage sequence. Visual stimuli were presented by means of a mirror mounted on the MRI head coil, and responses were acquired via an MRI-safe button response pad (Neuroscan).

Computational Modeling. To characterize the relative contributions of reinforcement (RL) and belief-based learning to behavior, we considered three different models of learning: reinforcement learning, belief-based learning, and their hybrid, experience-weighted attraction (EWA). We first describe the hybrid model because it contains RL and belief learning models as special cases (2). First, denote s_i^k as strategy k for player i, $s_i(t)$ is the chosen strategy by player i at period t, and $s_{-i}(t)$ is the chosen strategy of the opponent at period t. Player i's expected reward, $V_i^k(t)$, for playing strategy s_i^k in period t is governed by three parameters and updates according to the following:

$$V_i^k(t) = \begin{cases} \dfrac{\phi_i \cdot N(t-1) \cdot V_i^k(t-1) + \pi_i(s_i^k, s_{-i}(t))}{N(t)}, & \text{if } s_i^k = s_i(t) \\ \dfrac{\phi_i \cdot N(t-1) \cdot V_i^k(t-1) + \delta_i \cdot \pi_i(s_i^k, s_{-i}(t))}{N(t)}, & \text{if } s_i^k \neq s_i(t), \end{cases} \qquad \textbf{[S1]}$$

where parameter ϕ_i and function $N(t) = \rho_i \cdot N(t-1) + 1$ capture different aspects of the depreciation of V_i^k (t). For example, if the player believes his opponent is a fast adaptor, he will have a small ϕ_i that depreciates past values faster. In contrast, ρ_i is the discount rate for the strength of past experience $N(t)$, and controls the influence of the out-of-game prior beliefs. If ρ_i is large, the out-of-game prior beliefs will wear off quickly. The third and most important parameter for our study, δ_i, is the weight between foregone payoffs and actual payoffs when updating values, and reflects one of the key insights of the hybrid model that belief learning is equivalent to a model whereby actions are reinforced by foregone payoffs in addition to received payoffs as in RL models. Thus, δ_i can be interpreted as a psychological inclination toward belief learning (2). That is, the hybrid model reduces to the RL model when $\delta_i = 0$, and the belief learning model when $\delta_i = 1$.

In belief learning, we also impose the restriction that the initial attractions are expected payoffs given some underlying probabilistic belief inference of the subject, that is, $V_i^k(0) = \sum_m q_{-i}^m(0) \times \pi_i(s_i^k, s_{-i}^m)$, where $q_{-i}^m(0)$ is player i's initial belief about the likelihood of his opponent adopting s_{-i}^m. Hence, $q_{-i}^m(0) \geq 0$ and $\sum_m q_{-i}^m(0) = 1$. The restriction ensures that in all of the trials that follow the belief learners update a probabilistic belief inference regarding the next move of the opponents rather than an unconstrained vector of fictive errors defined as the discrepancy between forgone payoffs and previous attraction values.

Behavioral Data Analysis. To calibrate the models given behavior of the subjects in the game, we estimated parameters of each model by using responses of subjects by maximizing the logistic log likelihood of the model predictions. To convert values into choices, we used a logit or softmax function to calculate the probability of player i playing strategy k in the next round, $p_i^k(t+1) = e^{\lambda_i \cdot V_i^k(t)} / \sum_{l=1}^{L} e^{\lambda_i \cdot V_i^l(t)}$, where λ_i is a measurement of sensitivity of subjects to difference in expected reward associated with the different actions.

Using these choice probabilities, we performed maximum likelihood estimation with a grid search over a large range of values for all free parameters in all estimations, because the likelihood function is not globally concave. Both pooled and individual-level estimations were performed. For pooled estimation, we aggregated observations conditional on the roles of the subjects and then fit the choice data by maximizing the log likelihood of the observed choices over rounds for subject $_i$. That is, $\sum_i \sum_t \log(p_i^{s_i(t)}(t))$. Although using pooled estimates is more robust in general, it removes the possible individual variation in learning, and will bias estimates due to heterogeneity (3). Therefore, we also performed estimation at the individual level. The primary challenge of individual estimation is the relatively small sample size compared with the number of free parameters. We approached this problem with two methods combined: (*i*) estimating a common set of initial attractions shared by all subjects with the same role, from the pooled first period of data, conditional on the role of the subject and (*ii*) self-tuning estimation as introduced in Ho et al. (4). As a robustness check, we also conducted individual level estimation with partially joint estimates across different roles, assuming each subject shares a subset of learning parameters (e.g., the decay rate of the initial belief) regardless of her role in the game. We found that the estimates to be robust across these different estimation strategies.

Conversion to Temporal Difference Form. To derive trial-by-trial predictors for use in neuroimaging analysis, we converted the respective models above to a TD form whereby learning results from updating reward predictions through a prediction error. Choice probabilities and prediction errors on each trial were then generated by using the best-fit parameters derived from the behavioral data estimation. That is, we separated player *i*'s expected reward, $V_i^k(t)$, for playing strategy s_i^k in period t into a reward prediction $V_i^k(t-1)$, and the prediction error that is the difference between the expected reward and obtained (foregone) reward $\pi_i(s_i^k, s_{-i}(t))$. In the hybrid model, the expected reward thus evolves according to:

$$V_{i,k}^{EWA}(t) = \begin{cases} V_{i,k}^{EWA}(t-1) + \dfrac{1}{N(t)}\left\{\pi_i(s_{i,k}, s_{-i}(t)) - V_{i,k}^{EWA}(t-1)\right\} & \text{if } s_{i,k} = s_i(t) \\ \underbrace{V_{i,k}^{EWA}(t-1)}_{\text{Reward Prediction}} + \dfrac{1}{N(t)}\underbrace{\left\{\delta_i \cdot \pi_i(s_{i,k}, s_{-i}(t)) - V_{i,k}^{EWA}(t-1)\right\}}_{\text{Prediction Error}} & \text{if } s_{i,k} \neq s_i(t). \end{cases} \quad \textbf{[S2]}$$

In contrast, RL updates by reinforcing only the chosen strategy, whereas belief learning updates by reinforcing all available strategies proportional to the possible rewards:

$$\text{RL}: V_{i,k}^{RL}(t) = \begin{cases} V_{i,k}^{RL}(t-1) + (1-\phi_i)\left\{\dfrac{1}{1-\phi_i}\pi(s_{i,k}, s_{-i}(t)) - V_{i,k}^{RL}(t-1)\right\} & \text{if } s_{i,k} = s_i(t) \\ \phi_i \cdot V_{i,k}^{RL}(t-1) & \text{if } s_{i,k} \neq s_i(t) \end{cases} \quad \textbf{[S3]}$$

$$\text{BB}: V_{i,k}^{BB}(t) = V_{i,k}^{BB}(t-1) + \frac{1}{N(t)}\left\{\pi(s_{i,k}, s_{-i}(t)) - V_{i,k}^{BB}(t-1)\right\}, \forall\, s_{i,k} \quad \textbf{[S4]}$$

fMRI Data Analysis. Image analysis was performed by using *SPM2* (Wellcome Department of Imaging Neuroscience, Institute of Neurology, London). Preprocessing included, in order: slice time correction (centered at *TR*/2), motion correction, coregistration, spatial normalization to the Montreal Neurological Institute (MNI) template, and spatial smoothing using an 8-mm Gaussian kernel (5). All images were also high-pass filtered in the temporal domain (width 128s) and autocorrelation of the hemodynamic responses was modeled as an *AR*(1) process.

Analyses of fMRI time series were done by using standard random effects models (6), with reward prediction and prediction error values generated from the respective computational models calibrated on choices of subjects at the individual level. An event-related design was used where regressors were included for the decision and feedback events of the trials (Fig. 1). That is, for each subject, we constructed a (first level) general linear model (GLM) consisting of two events: an event at the time of decision, and one at the time of feedback. Regressors were constructed by using the trial-by-trial outputs from the TD form of the best-fitting individual parameter estimates. The decision event was associated with choice probabilities, which can be regarded as relative reward predictions controlled for time influence. The feedback event was associated with prediction errors for chosen actions. All analyses were performed on the feedback event data, except the expected reward region analysis (Fig. S3). The first eight rounds were excluded from the GLM analysis to allow initial values to stabilize. Regressors were convolved with the canonical hemodynamic response function and entered into a regression analysis against each subject's BOLD response data. The regression fits of each computational signal from each individual subject were then summed across their roles and then taken into random-effects group analysis.

SI Results

Comparison of Behavior Across Experimental Protocols. To measure the effectiveness of this protocol, we compared both aggregate choices and model estimates among (*i*) our neuroimaging subjects, (*ii*) our behavioral subjects, and (*iii*) Rapoport and Amaldoss's original experiment (7) (Table S1). Proportions of choices are similar, as are parameter estimates across the three different datasets. We found no evidence that the pool player protocol systematically affected behavior of players.

To further check the robustness of our pool player protocol, we compared behaviors in our strategic setting versus those in a matching but nonstrategic reward task. In the reward treatment, we replaced the human pool players with a computer algorithm. In contrast to the strategic treatment, subjects in the reward treatment were told to exceed a random hurdle determined by the computer to win the prize. Subjects were informed that they are playing against a computer algorithm. All other aspects of the instructions remained identical. In terms of the game display, the only difference was that in the reward treatment, the word "Opponent" was replaced with the word "Hurdle".

We found that learning in a reward setting is primarily RL-based. Using model-based estimates, we found that the hybrid δ parameter was significantly greater in the strategic treatment than the reward treatment ($P < 0.01$, two tailed). Visually, this difference can be illustrated through the transition matrices of the choices of players (Fig. S1). These matrices show how players switched their choices from one trial to the next and are generalizations of more traditional switch/stay measures (2). The diagonal elements indicate choices in which subjects stayed, whereas off-diagonals indicate switches.

The most striking features are the similarities between the transition matrices of the strategic treatment and the belief learning simulation (Fig. S1 *A* and *C*) and between reward treatment and RL simulation (Fig. S1 *B* and *D*). In particular, whereas players in the strategic treatment switched quite often, players in the reward treatment repeatedly played the same strategies, rarely switching between strategies from trial to trial. This behavior is apparent in that most of the mass of the transition matrix for the RL treatment and simulation is located along the diagonal (indicating stay trials) at investments of 1, 3, and 5 (Fig. S1*B*). At the aggregate level, the switch rate in the strategic treatment was 0.56 (exactly that of the Nash equilibrium prediction) versus 0.32 for the reward treatment. This finding is thus is consistent with the hypothesis that learning in the reward treatment is subserved primarily by reinforcement learning, which adapts more slowly.

Correlation of RL and Belief-Based Prediction Errors. Table S3 shows the correlation between the prediction errors associated with the three models under consideration. Crucially, we find that the correlation between RL and belief prediction errors is low (Pearson $\rho = 0.28$). The statistical separation between the model-generated learning signals indicates the potential to disentangle the unique contributions of the different types of learning signals. The correlation of reinforcement and belief-based learning with the hybrid model is not surprising, given the reinforcement and belief learning are nested models.

Orthogonality Tests on Robustness of Brain Activations. Although the correlation between RL and belief prediction errors was low, we nevertheless sought to investigate whether our reinforcement and belief prediction errors are robust to orthogonalization of the regressors. We verified that activations in response to RL and belief prediction errors remain after they are orthogonalized against each other (Fig. S2). The procedure is same as those described in ref. 8.

Expected Reward Regions. We found activity in ventromedial prefrontal cortex, extending to rACC and medial orbitofrontal cortex, to be correlated with the relative expected reward value of the chosen action (Fig. S3). The relative expected reward is defined as the probability generated from the different models for the chosen action at the time of response on a given trial. We used this notion to remove the possible time trend in the absolute expected reward values. This result is consistent with existing evidence on the role of orbital and adjacent medial prefrontal cortex in encoding predictions of future reward (9, 10).

1. Deichmann R, Gottfried JA, Hutton C, Turner R (2003) Optimized EPI for fMRI studies of the orbitofrontal cortex. *Neuroimage* 19:430–441.
2. Camerer CF, Ho T (1999) Experience-weighted attraction learning in games: A unifying approach. *Econometrica* 67:827–874.
3. Wilcox NT (2006) Theories of learning in games and heterogeneity bias. *Econometrica* 74:1271–1292.
4. Ho T, Camerer C, Chong J (2007) Self-tuning experience weighted attraction learning in games. *J Econ Theory* 133:177–198.
5. Friston KJ, et al. (1995) Statistical parametric maps in functional brain imaging: A general linear approach. *Hum Brain Mapp* 2:189–210.
6. Friston KJ, Stephan KE, Lund TE, Morcom A, Kiebel S (2005) Mixed-effects and fMRI studies. *Neuroimage* 24:244–252.
7. Rapoport A, Amaldoss W (2000) Mixed strategies and iterative elimination of strongly dominated strategies: An experimental investigation of states of knowledge. *J Econ Behav Organ* 42:483–521.
8. Lohrenz T, McCabe K, Camerer CF, Montague PR (2007) Neural signature of fictive learning signals in a sequential investment task. *Proc Natl Acad Sci USA* 104: 9493–9498.
9. Daw ND, O'Doherty JP, Dayan P, Seymour B, Dolan RJ (2006) Cortical substrates for exploratory decisions in humans. *Nature* 441:876–879.
10. O'Doherty JP, et al. (2004) Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science* 304:452–454.

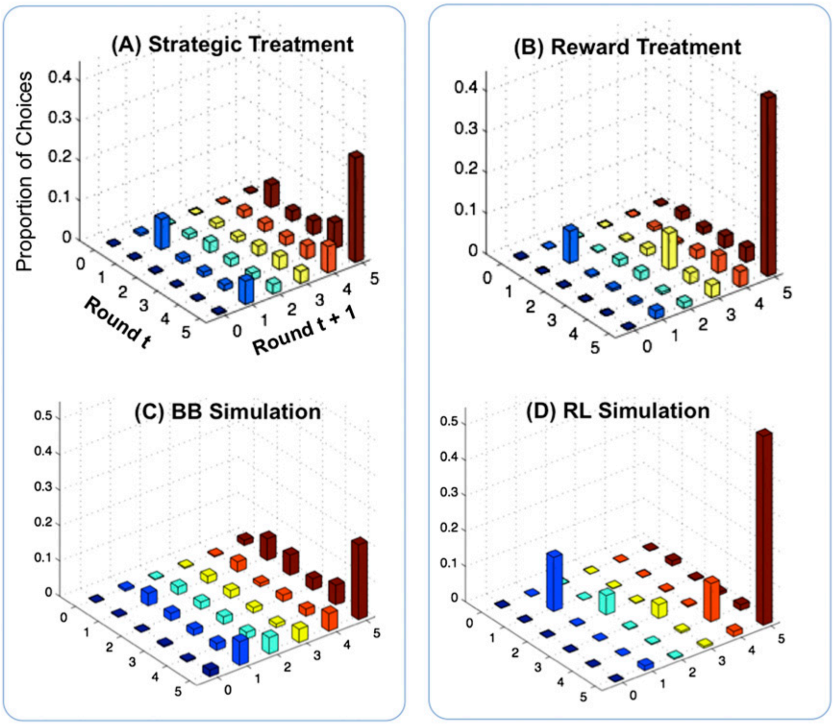

Fig. S1. Comparison of strategic and reward learning in Strong role. (*A* and *B*) Empirical frequency of transitions for strategic and reward treatments, respectively. (*C* and *D*) Transition matrices of simulations using belief and reinforcement learning models, respectively. Note behavior in strategic treatment is qualitatively more similar to the belief learning simulation, whereas reward treatment is more similar to the RL simulation.

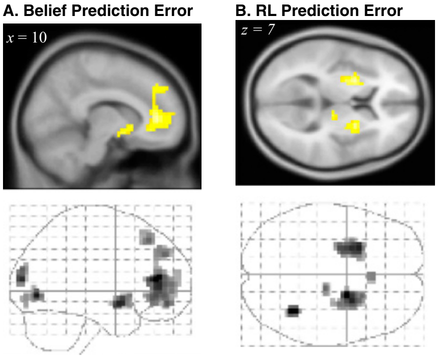

Fig. S2. Robustness check for orthogonalization between RL and belief learning prediction errors. (*A*) Belief learning prediction errors after orthogonalization against RL prediction errors. ($P < 0.001$, uncorrected, cluster size $k > 10$ voxels). (*B*) RL prediction errors after orthogonalization against belief learning prediction errors. ($P < 0.001$, uncorrected, cluster size $k > 10$ voxels).

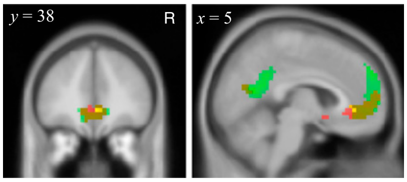

Fig. S3. Expected reward regions. Activity in ventromedial prefrontal cortex, extending to rACC and medial orbitofrontal cortex, is correlated with respect to relative expected reward value of the chosen action calculated under the hybrid (red), belief (yellow), and RL (green) models ($P < 0.005$ uncorrected, cluster size $k \geq 5$).

Table S1. Comparison of Nash equilibrium predictions and empirical distributions from (*i*) Rapoport and Amaldoss (1), (*ii*) our behavioral experiment, (*iii*) our neuroimaging experiment, and (*iv*) a reward learning control session

			Empirical distributions			
Role	Investment	Equilibrium prediction, %	Matrix form, %	Behavioral session, %	Neuroimaging session, %	Reward learning, %
Strong	0	0	1	0	1	1
	1	20	17	14	18	11
	2	0	5	6	10	6
	3	20	9	13	11	16
	4	0	13	25	16	11
	5	60	55	43	45	54
Weak	0	60	55	49	49	30
	1	0	3	3	4	12
	2	20	6	10	7	18
	3	0	14	10	14	8
	4	20	22	28	27	32

Empirical distribution is proportion of all players' choices over all rounds.

Table S2. Median individual level estimates

Model	δ	ϕ	λ
Reinforcement	0*	0.94 (0.86, 0.96)	0.04 (0.02, 0.07)
Belief-based	1*	0.95 (0.83, 0.98)	0.60 (0.23, 2.11)
Hybrid	0.46 (0.29, 0.69)	0.71 (0.53, 0.81)	0.51 (0.32, 0.70)

Parentheses contain first and third quartile of empirical distribution.
*Parameters constrained by model.

Table S3. Correlation coefficient between the prediction errors from different learning models

Model	Reinforcement	Belief-based	Hybrid
Reinforcement	—	(0.16)	(0.10)
Belief-based	0.28	—	(0.18)
Hybrid	0.63	0.40	—

Parentheses contain SDs for the correlation coefficients.