How instructed knowledge modulates the neural systems of reward learning

Jian Li[a,b], Mauricio R. Delgado[c], and Elizabeth A. Phelps[a,b,1]

[a]Department of Psychology, [b]Center for Neural Science, New York University, New York, NY 10003; and [c]Department of Psychology, Rutgers University, Newark, NJ 07102

Edited by Edward E. Smith, Columbia University, New York, NY, and approved November 19, 2010 (received for review October 5, 2010)

Recent research in neuroeconomics has demonstrated that the reinforcement learning model of reward learning captures the patterns of both behavioral performance and neural responses during a range of economic decision-making tasks. However, this powerful theoretical model has its limits. Trial-and-error is only one of the means by which individuals can learn the value associated with different decision options. Humans have also developed efficient, symbolic means of communication for learning without the necessity for committing multiple errors across trials. In the present study, we observed that instructed knowledge of cue-reward probabilities improves behavioral performance and diminishes reinforcement learning-related blood-oxygen level-dependent (BOLD) responses to feedback in the nucleus accumbens, ventromedial prefrontal cortex, and hippocampal complex. The decrease in BOLD responses in these brain regions to reward-feedback signals was functionally correlated with activation of the dorsolateral prefrontal cortex (DLPFC). These results suggest that when learning action values, participants use the DLPFC to dynamically adjust outcome responses in valuation regions depending on the usefulness of action-outcome information.

functional MRI | striatum | instruction | computational modeling | prediction error

Maximizing reward obtained over time can be a daunting challenge to any organism (1). Without concrete instruction, an animal can only develop and fine-tune its reward-harvesting strategy through trial and error. Reinforcement learning (RL) theory has formalized this intuition and associated prediction error to the phasic changes of activities in dopaminergic neurons that track the ongoing difference between experienced and expected reward (2). Under this framework, prediction error is thought to broadcast to valuation structures, such as the striatum and ventromedial prefrontal cortex (vmPFC), to direct learning and integrate with other streams of information to facilitate decision making (3–6).

Recent research in neuroeconomics has demonstrated that this theoretical model of trial-and-error reward learning captures the patterns of both behavioral performance and the blood-oxygen level-dependent (BOLD) signals during a range of economic decision-making tasks (3, 7–10), demonstrating important cross-species similarities in the mechanisms of reward learning (11). Because of this finding, RL models have become a primary means to characterize neural responses in neuroeconomics. However, this powerful theoretical model has its limits. Trial and error is only one of the means by which individuals can learn the value associated with different decision options. Humans have also developed efficient, symbolic means of communication, namely language, that allow the social communication of information about value without the necessity for committing multiple errors across trials to learn. Little is known about how this explicit, symbolic knowledge can infiltrate the valuation structures mentioned above and exert its influence on action selection, and how the brain's embodiment of the RL algorithm differs in the face of instructed knowledge.

To address these questions, we used functional MRI (fMRI) together with a probabilistic reward task (Fig. 1) to assess the relative contributions of trial-and-error feedback and instructed knowledge on choice selection (12, 13). We designed an experiment with two sessions. In the "feedback" session, participants' choices were only based on the win/loss feedback, and in the "instructed" session participants could also incorporate the correct cue-reward probability information provided by experimenter to guide choice behavior. We hypothesize that: (*i*) RL is a robust algorithm to explain and predict choice behaviors and BOLD responses in an environment where trial-and-error feedback is the only information to guide learning and influence choices (13–17), and (*ii*) when instructed knowledge about reward probabilities is also available, participants use this extra information to achieve better performance by modulating the degree to which RL algorithms are involved. We also explored which brain systems may influence the implementation of instructed knowledge by modulating the patterns of BOLD responses in brain areas typically implicated in RL, valuation and choice selection (13, 18–26).

Results

Behavioral Results. For both sessions, the participants' frequency of win trials varied across four different probability conditions ($F_{3,76} > 90$, $P < 0.0001$). A repeated measure two-way ANOVA was performed using probability condition (25, 50, 75, and 100%) and session condition (feedback and instructed) as within-participant factors. As expected, participants showed better performance, as indicated by the frequency of win trials, in the instructed session ($F_{1,152} = 5.8$, $P < 0.02$) (Fig. 2*A*). Post hoc *t* tests showed that the differential performance existed in the 25, 75, and 100% probability conditions ($P < 0.05$), but not the 50% condition ($P = 0.16$) (Fig. 2*A*). We then examined how participants' choices were influenced by the most recent outcome they received upon their subsequent choices. We calculated how likely participants would stay with their previous choice when the outcome of the previous choice was positive (win) or negative (loss) in both sessions (Fig. 2*B*). Using previous trial outcome (win or loss) and session condition (feedback or instructed) as two factors, a two-way ANOVA revealed that there is a main effect for outcome type. Participants were more likely to follow their previous choice when the outcome was positive ($F_{1,76} = 26.9$, $P < 0.001$). Further post hoc analysis showed participants in the feedback session were more influenced in their subsequent choice action by positive outcomes than participants in the instructed session ($P < 0.05$). A similar trend was observed for negative outcomes, but it was not significant ($P = 0.18$) (Fig. 2*B*). Overall, the loss trials were less than 30% of all of the trials, resulting in diminished statistical power for analyses of loss, relative to win, trials.

We fitted a Q-learning model to participants' choice behavior in both the feedback and instructed sessions using the maximum

Author contributions: J.L., M.R.D., and E.A.P. designed research; J.L. performed research; J.L. contributed new reagents/analytic tools; J.L. analyzed data; and J.L., M.R.D., and E.A.P. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

[1]To whom correspondence should be addressed. E-mail: liz.phelps@nyu.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1014938108/-/DCSupplemental.

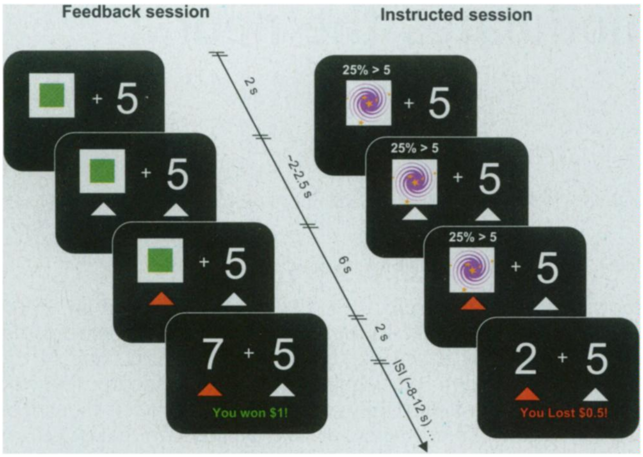

Fig. 1. Experimental design. In feedback session, the number 5 and a specific visual cue were displayed on the screen. In the instructed session, additional probability information was displayed on top of the visual cue.

likelihood estimation. Different prominent RL models were fitted to the participants' behavioral data to determine the optimal model. We considered popular models: a RL model with a single learning rate for both positive and negative prediction errors (PEs) (δ_+ and δ_-), and a RL model with different learning rates for both positive and negative PEs (δ_+ and δ_-). In the instructed session, we also included a RL model that assigns a "confirmation bias" to outcomes that match the instructions (27, 28) (*SI Appendix*, Tables S1 and S2). In the feedback session, a simple RL model with a single learning rate (α) for both positive and negative PEs (δ_+ and δ_-) tended to fit participants' behavior better. However, a Q-learning model with two different learning rates (α_+ and α_-) for positive and negative PEs (δ_+ and δ_-) best explained participants' behavior in the instructed session (implementation of model fitting is detailed in *SI Appendix*). The McFadden's pseudo R-square was 0.50 for the feedback session and 0.61 for the instructed session (*SI Appendix*, Table S1). A single learning rate of 0.24 was estimated for the feedback session; however, the learning rates associated with positive PE (α_+) was 0.05 and was 0 with negative PE (α_-) in the instructed session (*SI Appendix*, Tables S1 and S2). The significant difference of learning rates indicates that the PEs were not as efficiently incorporated to the updating of action value in the instructed session, especially when the outcome was worse than participants' expectation. These results suggest that participants' actions were simply governed by the a priori action value instructed by experimenter, as indicated by the initial Q value associated with different stimuli ($\alpha_- = 0$) (*SI Appendix*, Table S2).

Functional MRI Results. ***RL model predicts BOLD signals in the feedback session.*** Our behavioral results suggest that a RL model captures participants' performance in the feedback session, but does not adequately describe learning in the instructed session. To explore if a similar pattern was reflected in the patterns of BOLD responses, we constructed a general linear model (GLM) with the PE regressors generated from the best fitting Q-learning models for both sessions (*SI Appendix*, Tables S1 and S2) and investigated the neural correlates of PE in the feedback session and the instructed session (Fig. 3*A*). Ventral striatum BOLD response was significantly correlated with PE signals in the feedback session [$P < 0.05$, corrected, peak Montreal Neurological Institute (MNI coordinate) (–27 3 0), $z = 3.59$] (Fig. 3*A* and *SI Appendix*, Table S2). There was no such correlation observed in the instructed session, even under a more relaxed threshold ($P < 0.01$, uncorrected). A direct comparison between the BOLD responses that correlated with PEs in the feedback and instructed sessions further confirmed the differential involvement of striatum in encoding PEs in both sessions (*SI Appendix*, Fig. S1). Because PE and monetary outcome often tend to correlate with each other, PE regression analyses were performed by including the monetary outcome regressor in the GLM for both sessions to separate PE-related BOLD responses from the outcome-related ones.

As a learning signal, PE has its unique activity pattern. At the beginning of the learning phase, PE signals tend to respond to the onset of the outcome delivery, but as learning progresses PE signal shifts toward the onset of the cue/decision accompanied by a diminished response to the actual outcome. To further test that BOLD response in the striatum indeed encodes PEs in the feedback session, we conducted an independent two-way ANOVA to identify brain regions whose activities were modulated by the interaction between the session (instructed vs. feedback) and learning phase (early vs. late learning). We hypothesized that if a RL mechanism was engaged differentially between the instructed and the feedback session, we should observe an interaction between the session and learning phase factors in a two-way ANOVA. Indeed, this analysis yielded similar brain regions as the PE regression analysis [$P < 0.05$, small volume corrected for 343 surrounding voxels, peak MNI coordinate (–15 –6 0), $F_{1,304} = 15.55$, $z = 3.72$] (Fig. 3*B*). A further region of interest (ROI) time-series analysis in the overlapping region of activation in the ventral striatum (Fig. 3*C*) revealed a pattern of BOLD response consistent with a PE learning signal in the feedback session. In early trials, striatum activation peaked at the onset of outcome. As learning progressed, this peak activation shifted toward the decision onset

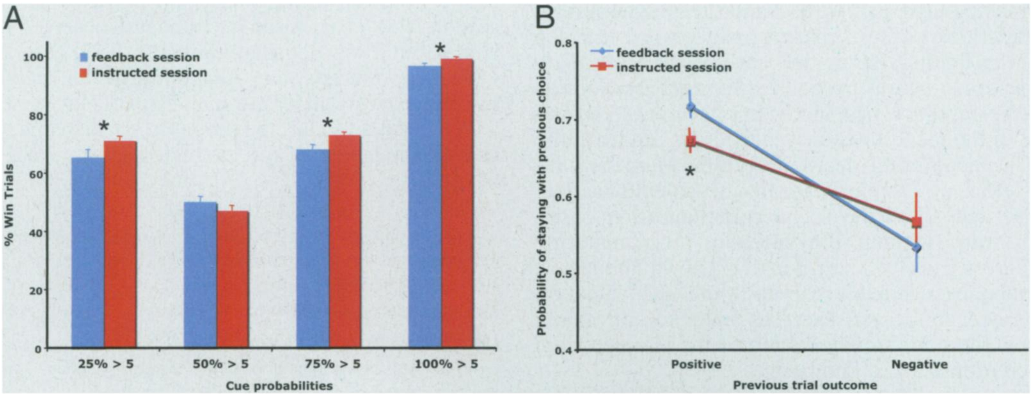

Fig. 2. Behavioral results for both sessions. (*A*) Percentage of win trials (±SEM) for the different visual cue probabilities for both sessions. (*B*) The probability of staying with the previous choice (±SEM) given its outcome (positive or negative) for both sessions. (*, significant difference between sessions, $P < 0.05$).

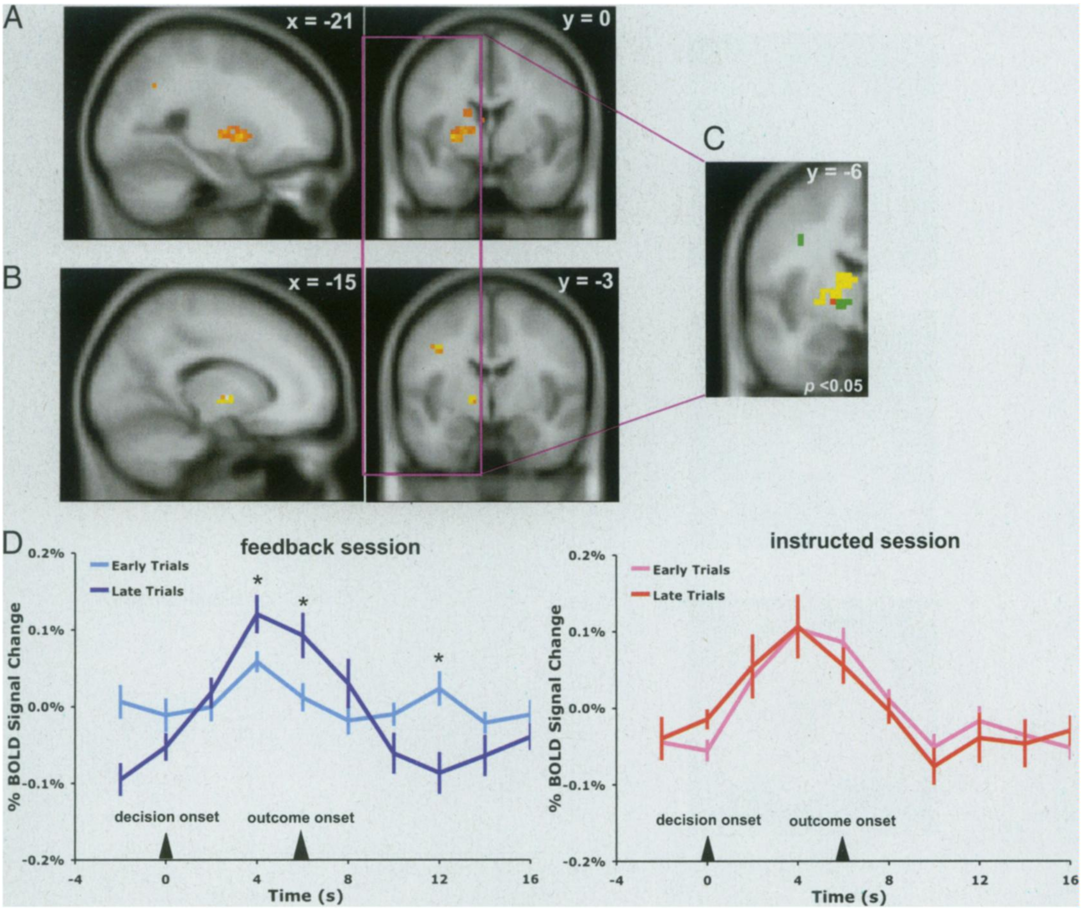

Fig. 3. BOLD responses for prediction errors in both sessions. (*A*) Activity of the striatum showed significant correlation to the PE signal in the feedback session ($P < 0.05$, corrected). Such correlations were not observed in the above structures in the instructed session ($P < 0.01$, uncorrected). (*B*) A two-way ANOVA showed an interaction between session (feedback and instructed) and learning phase (early and late) in the left striatum. (*C*) Striatal activation identified in the PE (*A*, yellow) and session × learning phase interaction (*B*, green) analyses, and the overlapping region (red). (*D*) BOLD response patterns in the overlapping region for the early and late phases of learning in the feedback and instructed sessions (*, time points with significantly different BOLD responses between early and late learning phases, $P < 0.05$; ± SEM).

(Fig. 3*D*, *Left*). However, this characteristic PE response pattern was absent in the instructed session (Fig. 3*D*, *Right*).

Reduced BOLD responses to outcomes in the instructed session. Inspired by the results that participants' choices are differentially influenced by previous trial outcomes (Fig. 2*B*), we examined participants' BOLD responses when participants processed monetary outcome (win or loss) in both sessions. From a general contrast of win over loss at the onset of outcome revelation across both sessions, we found significant activation in the nucleus accumbens (NAc) [peak MNI coordinate (−2 12 −10), z = 6.12] and vmPFC [peak MNI coordinate (−4 42 −10), z = 5.94; P < 0.05 corrected] (Fig. 4*A*), regions previously linked to the brain's reward valuation system (25, 29–34) (*SI Appendix*, Table S3). In addition, we observed bilateral activation in the hippocampal complex [peak MNI coordinates (−18 −18 −20), z = 4.31 and (28 −18 −20), z = 3.61], which was centered on the perirhinal cortex, a region that has been implicated in processing item-reward associations (35–37) (Fig. 4*A*). Similar outcome-related activation patterns were also observed by including the prediction error regressor in the GLM.

ROI analyses of these brain regions showed that overall BOLD responses to outcomes (win minus loss) were smaller in the instructed session than the feedback session. Examining win and loss trials independently revealed diminished activation to monetary gains in the instructed session in all three regions (P < 0.05 at the peaks of activation). Although a similar pattern was observed for loss trails, no significant differences were observed for loss evoked responses between the two sessions, perhaps because of the diminished statistical power resulting from fewer overall loss trials (Fig. 4*B*).

Higher dorsolateral prefrontal cortex activity paralleled better performance in the instructed session. RL model fitting of the behavioral data suggested that participants were less influenced by monetary outcomes in the instructed session, most likely because of the strong a priori instructed knowledge of the cue-reward probabilities. Accordingly, participants achieved better performance in the instructed session. This reliance on instructed knowledge reduced BOLD responses in regions implicated in reward learning, suggesting that instructed knowledge enables the brain to diminish the impact of outcome feedback on decision making. If this process is the case, there should also be a corresponding increase in activation in brain regions that mediate the implementation of instructed knowledge. To determine which brain regions may enable the effects of instructed knowledge on trial-and-error reward learning tasks, we conducted an exploratory analysis to locate brain areas where activation to monetary outcomes was greater in the instructed relative to feedback session. We focused on win trials because our previous analyses found significantly diminished BOLD responses to wins in reward learning (NAc and hippocampal complex) and valuation (vmPFC) regions in the instructed session. This analysis revealed the left dorsolateral prefrontal cortex (DLPFC) [P < 0.05 corrected, peak MNI coordinate: (−48 24 33), z = 3.98]

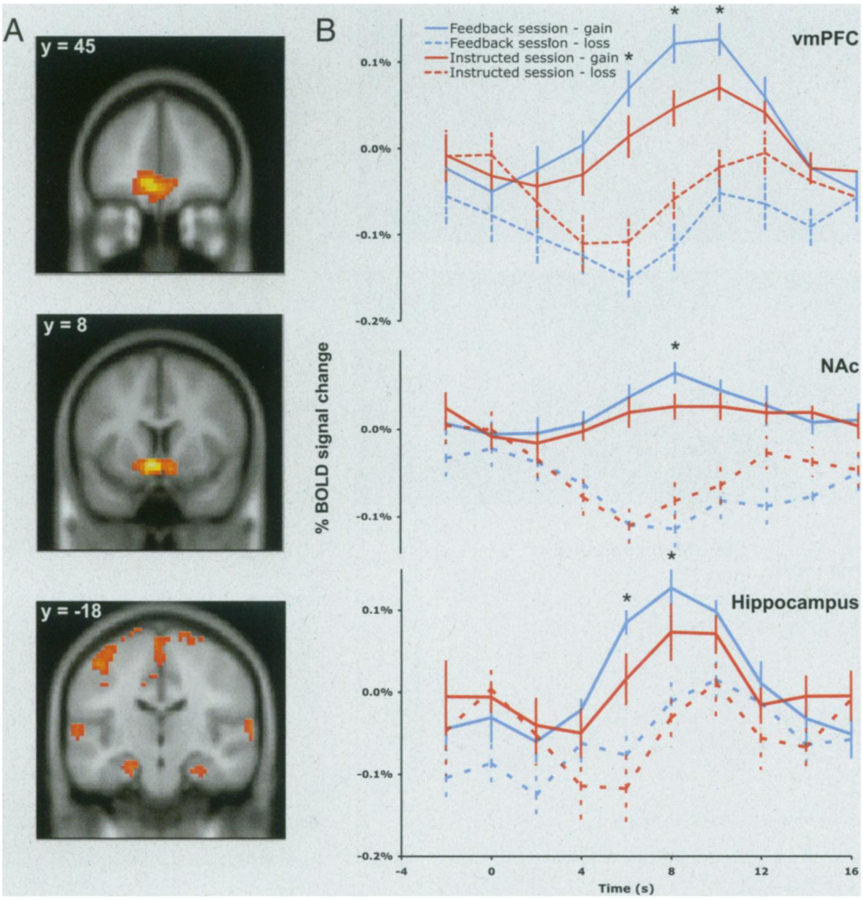

Fig. 4. BOLD responses discriminating win and loss for both sessions. (*A*) A whole-brain analysis revealed greater activation in the NAc, vmPFC, and bilateral hippocampal complex for win than loss trials across both sessions ($P < 0.05$, corrected). (*B*) BOLD time course of activation in the NAc, vmPFC, and bilateral hippocampal complex for win and loss trials in the feedback and instructed sessions (*, significant difference of time points near activation peaks, $P < 0.05$; ± SEM).

showed a greater BOLD response to win outcomes during the instructed session (Fig. 5*A* and *SI Appendix*, Table S5).

Functional connectivity between DLPFC and reward-related brain structures. The DLPFC has previously been implicated in decision-making and emotion regulation tasks that require the top-down modulation of valuation regions (25, 26, 34). To determine if the left DLPFC acted as a cognitive modulator of reward learning regions in the presence of instructed knowledge, we conducted a psychophysiological interaction (PPI) analysis using the peak voxels in the left DLPFC (Fig. 5*A*) as the seed region, and tested which brain areas showed significant functional connectivity in the win trials vs. loss trials. We found an inverse, win-trial specific functional connectivity between the DLPFC and the NAc [peak MNI coordinate (−3 6 −12), $z = 3.17$], vmPFC [peak MNI coordinate (−6 48 −18), $z = 4.86$], and left parahippocampal gyrus [peak MNI coordinate (−36 −24 −18), $z = 3.77$] only in the instructed session ($P < 0.05$ corrected) (Fig. 5*B* and *SI Appendix*, Table S6). Similar results were obtained by directly comparing the functional connectivity of the DLPFC and these reward-related brain areas in the feedback and instructed session (*SI Appendix*, Fig. S2). This result is particularly interesting because the valuation structures, whose BOLD responses are negatively correlated with the left DLPFC when reliable instructed knowledge is available to guide choices (vmPFC, NAc, and hippocampal complex), overlap with those regions showing diminished response to reward outcomes in the instructed session (Figs. 4 and 5).

Discussion

Optimal decision making requires the brain to dynamically allocate control among different types of information for action selection (17, 38, 39). When feedback is the only source of information, choice-dependent outcomes can be evaluated and fed back to valuation systems to provide a better approximation of action values and guide individuals toward choices that maximize accumulated rewards in the long run. The RL algorithm provides a formal framework to incorporate feedback information to facilitate learning and decision-making (1, 2, 40–43). Consistent with this previous research, in the feedback session of our task we fit a RL model to participants' behavioral data and located the neural basis of PE in the ventral striatum (Fig. 3 and *SI Appendix*, Table S1) using two independent approaches (Fig. 3 *A* and *B*). Additional ROI time-series analyses in the feedback session further revealed that striatal BOLD responses were sensitive to the onset of both decisions and outcomes early in learning, but migrated to the onset of the decision as learning progressed. This pattern is consistent with the unique characteristics of PE learning signals and is absent in the instructed session (Fig. 3*D*) (2).

In contrast, when correct instructed knowledge about the cue-reward probabilities was available, participants used this information to achieve better performance and the RL model was less successful in interpreting participants' behaviors and BOLD activation pattern (Fig. 3*D* and *SI Appendix*, Fig. S1). Previous research has formalized the intuition of instructional control and suggested a "confirmation bias" model to amplify the effect of positive PEs and diminish the effect of negative PEs when participants made choices based on instructed information (27, 28). We compared the performance of different RL models [including the confirmation-bias model suggested by Doll et al. (27)] and interestingly, the RL model with different learning rates (α_+ and α_-) for positive and negative PEs (δ_+ and δ_-) tended to fit par-

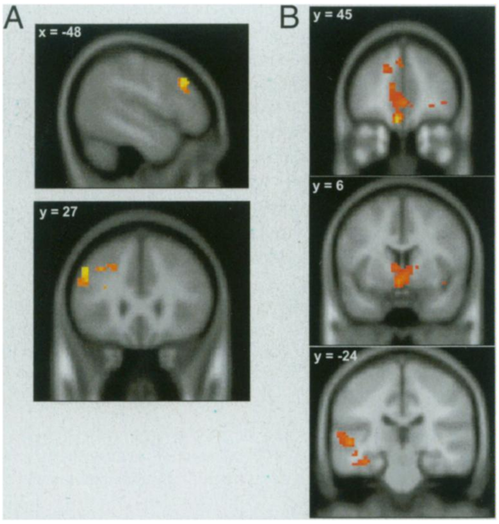

Fig. 5. Left DLPFC activity showed negative functional connectivity to brain structures related to reward valuation. (*A*) Left DLPFC showed relatively greater activation to monetary gains in the instructed than the feedback session ($P < 0.05$, corrected). (*B*) PPI analysis showing regions negatively correlated with the left DLPFC on win trials in the instructed session ($P < 0.05$, corrected) but not in the feedback session ($P < 0.01$, uncorrected) (*SI Appendix*, Fig. S2).

ticipants' behavior best in the instructed session (see *SI Appendix* for technical details). Using the PEs generated from the above best-fitting RL model, our fMRI analysis did not reveal a correlation between PE signal and striatal BOLD responses ($P < 0.01$, uncorrected) in the instructed session. Taken together, these results suggest that participants might rely less on PE signals for action-value updating when symbolic, instructed knowledge of the reward probabilities is available. Consistent with this hypothesis, both behavior (Fig. 2*B*) and BOLD responses were less influenced by outcomes in the instructed session. Indeed, we observed relatively smaller activations in brain areas (NAc, vmPFC, and hippocampal complex) typically associated with reward learning and valuation (11, 25, 30, 31, 33, 35–37, 44–50) when participants were rewarded for their choices in the instructed session (Fig. 4*B* and *SI Appendix*, Table S4). These findings suggest that the brain assigns less weight to actual outcomes when other sources of reliable information (instructed knowledge) about the cue-reward probability and optimal choice strategies are available.

The mechanism by which participants dynamically adjust their reliance on outcome information when symbolic knowledge of the reward probabilities is available was revealed in an exploratory analysis that showed higher left DLPFC activity when participants experienced monetary wins in the instructed relative to the feedback session (Fig. 5*A*). Importantly, PPI analysis using this DLPFC area as a seed region revealed negative functional connectivity between BOLD activities in the left DLPFC, and those in brain regions related to reward learning and valuation (NAc, vmPFC, and the hippocampal complex) among other brain areas (Fig. 5*B* and *SI Appendix*, Fig. S2 and Table S6). Interestingly, these regions overlapped remarkably well with those identified previously through the win-loss contrast (Figs. 4*A* and 5*B*). Thus, we propose a functional link between the DLPFC and an outcome-valuation learning system. This latter system is pivotal in providing correct value or "utility" information to facilitate learning based on monetary feedback, but appears to be less important when preexisting, symbolic knowledge to guide choices is available.

Taken together, these results suggest that when learning action values, the DLPFC tends to dynamically adjust outcome responses in reward-related brain regions depending on the usefulness of action-outcome information compared with explicit knowledge participants directly obtained from social communication. Although the current study demonstrates the importance of this DLPFC, reward-related structure circuitry for learning and reward processing, previous neuroeconomic research has outlined a similar circuitry across a range of decision-making tasks in which preexisting reward values that are represented in valuation regions can be modulated based on social processes (51, 52), goals (34), or other cognitive factors (52, 53). Although there have been suggestions that the DLPFC and reward-related regions represent independent systems in the brain competing with each other for the dominance of action selection (27, 28, 53, 54), our results are more consistent with a general role for the DLPFC in modulating the engagement of reward-related regions depending on the relative importance of the information during a learning paradigm.

Our findings also lend neurological evidence to support recent computational approaches to reconcile a broad range of literatures suggesting multiple representation systems in the brain for behavioral control. One such system deploys a model-free method and "learns putatively simpler quantities," such as policies that are sufficient to permit optimal performance through processing action outcomes. It is suggested this computation is carried out in the dorsolateral striatum. The other system, which employs the prefrontal cortex, adopts a model-based method to make use of available or learned rules and derives optimal choice through dynamic programming. The brain arbitrates between different representation systems according to the uncertainty estimated from each system (38, 39, 55). In our task, the state transition probabilities (reward probabilities) were more accurate in the instructed session (provided by the experimenter), thus the model-based approach would dominate participants' choice by recruiting the DLPFC to bias responses in the reward-valuation systems.

The brain's reward-learning circuitry as instantiated in the RL model is a phylogenetically old system for learning based on trial-and-error. When social structures and means of communication are more complex, these basic reward-learning processes may not be optimal to promote the best decision. Errors are costly and unnecessary when additional, symbolic information about the best decision is available. The current study demonstrates how the DLPFC interacts with the reward-learning circuitry to diminish the impact of actual trial-outcome information, presumably enabling symbolic knowledge of reward probability to guide choices. Our data add to the growing literature of interactions of different types of information to achieve optimal behavior in decision making and provide direct support to the computational theory that arbitrates between different representation systems by assigning control to the one that has less uncertainty of the correct action values.

Methods

Participants. Twenty participants were recruited and tested in compliance with the university committee on activities involving human subjects [University Committee on Activities Involving Human Subjects (UCAIHS)]. The experiment was approved by the UCAIHS at New York University and all subjects provided informed consent before the experiment. Of the 20 participants, 7 were male, 9 were non-Caucasian, and the group had an average age of 21.6 y (SD = 3.72).

Experimental Procedures. Each participant played two sessions of the task. One session was named the "feedback" session and the other session was titled the "instructed" session (Fig. 1). For both sessions, participants were told that they would see different visual cues which represent how likely the number underneath the cue would be greater or less than 5 (value of underlying number $\in \{1, 2, 3, 4, 6, 7, 8, 9\}$). The sequence of the two sessions was randomized across participants, so that 10 out of 20 participants experienced the feedback session first. In both sessions, four different visual cues representing different probabilities ($P \in \{25, 50, 75, 100\%\}$) of the number underneath the cue being greater than 5 were presented to participants. For both sessions, participants saw a cue next to the number 5 on each trial. Each cue was randomly presented 20 times for a total of 80 trials per session (see *SI Appendix* for details).

Functional MRI Image Acquisition. Scanning was performed on all 20 participants with a 3-T Siemens Allegra head-only scanner and a Siemens standard head coil at New York University's Center for Brain Imaging (see *SI Appendix* for details).

Behavioral Analysis. Participants' choice behaviors in both sessions were modeled by a simple RL algorithm (See *SI Appendix* for details). We tested our model against others suggested in the literature based on behavioral data with similar tasks (27, 28) using the Bayesian information criterion as a criterion for model selection. For the feedback session, the simple RL with one learning rate (α) for both positive and negative prediction errors fits participants' behavior better. However, RL with different learning rates (α_+ and α_-) for positive and negative (δ_+ and δ_-) PEs fits participants' choices the best in the instructed session (see *SI Appendix* for details).

Imaging Analysis. We first regressed PEs that were generated for both the feedback and instructed sessions using the best-fitting parameters to the whole-brain BOLD signals at the revelation of monetary outcome to identify the brain areas whose activities were correlated with the calculation of PE. Monetary outcomes were also included as dummy regressors to account for the effect of the magnitude of the reward value.

Repeated-measures two-way ANOVA was performed on the functional imaging data with two factors (session and learning phase) at the onset of feedback. The finite impulse response from time 0 to ~12 s (TR0 to ~TR6) was generated by resampling the BOLD time series of each voxel in the brain and averaging across 40 trials each for the early and late learning phases in both sessions. Because canonical hemodynamic response function typically peaks at 6 to ~8 s after the stimulus onset, the two-way ANOVA was performed on both TR3 (6 s) and TR4 (8 s). These whole-brain analyses were performed on each voxel to identify brain regions that showed a significant interaction effect with time (i.e., early vs. late learning) and session (i.e., feedback vs. instructed session).

Finally, we conducted a PPI analysis to investigate the connectivity between brain regions that may modulate the impact of instructed knowledge on RL learning signals (see *SI Appendix* for technical details).

ACKNOWLEDGMENTS. We thank K. Sanzenbach and the Center for Brain Imaging at New York University for technical assistance. This study was funded by a James S. McDonnell Foundation grant and National Institute of Mental Health Grants MH 080756 (to E.A.P.) and MH 084081 (to M.R.D.).

1. Sutton RS, Barto AG (1998) *Reinforcement Learning: An Introduction* (MIT Press, Cambridge, Mass.), p 322.
2. Schultz W, Dayan P, Montague PR (1997) A neural substrate of prediction and reward. *Science* 275:1593–1599.
3. Montague PR, King-Casas B, Cohen JD (2006) Imaging valuation models in human choice. *Annu Rev Neurosci* 29:417–448.
4. Tanaka SC, et al. (2006) Brain mechanism of reward prediction under predictable and unpredictable environmental dynamics. *Neural Netw* 19:1233–1241.
5. Padoa-Schioppa C, Assad JA (2006) Neurons in the orbitofrontal cortex encode economic value. *Nature* 441:223–226.
6. Rudebeck PH, et al. (2008) Frontal cortex subregions play distinct roles in choices between actions and stimuli. *J Neurosci* 28:13775–13785.
7. McClure SM, Berns GS, Montague PR (2003) Temporal prediction errors in a passive learning task activate human striatum. *Neuron* 38:339–346.
8. King-Casas B, et al. (2005) Getting to know you: Reputation and trust in a two-person economic exchange. *Science* 308(5718):78–83.
9. Rangel A, Camerer C, Montague PR (2008) A framework for studying the neurobiology of value-based decision making. *Nat Rev Neurosci* 9:545–556.
10. Hare TA, O'Doherty J, Camerer CF, Schultz W, Rangel A (2008) Dissociating the role of the orbitofrontal cortex and the striatum in the computation of goal values and prediction errors. *J Neurosci* 28:5623–5630.
11. Breiter HC, Aharon I, Kahneman D, Dale A, Shizgal P (2001) Functional imaging of neural responses to expectancy and experience of monetary gains and losses. *Neuron* 30:619–639.
12. Delgado MR, Miller MM, Inati S, Phelps EA (2005) An fMRI study of reward-related probability learning. *Neuroimage* 24:862–873.
13. Burke CJ, Tobler PN, Baddeley M, Schultz W (2010) Neural mechanisms of observational learning. *Proc Natl Acad Sci USA* 107:14431–14436.
14. Kaelbling LP, Littman ML, Moore AW (1996) Reinforcement learning: A survey. *J Artif Intell Res* 4:237–285.
15. Bogacz R, McClure SM, Li J, Cohen JD, Montague PR (2007) Short-term memory traces for action bias in human reinforcement learning. *Brain Res* 1153:111–121.
16. Schönberg T, Daw ND, Joel D, O'Doherty JP (2007) Reinforcement learning signals in the human striatum distinguish learners from nonlearners during reward-based decision making. *J Neurosci* 27:12860–12867.
17. Niv Y (2009) Reinforcement learning in the brain. *J Math Psychol* 53(3):139–154.
18. Miller EK, Cohen JD (2001) An integrative theory of prefrontal cortex function. *Annu Rev Neurosci* 24:167–202.
19. Ochsner KN, Gross JJ (2005) The cognitive control of emotion. *Trends Cogn Sci* 9: 242–249.
20. LaBar KS, Cabeza R (2006) Cognitive neuroscience of emotional memory. *Nat Rev Neurosci* 7:54–64.
21. Li J, McClure SM, King-Casas B, Montague PR (2006) Policy adjustment in a dynamic economic game. *PLoS ONE* 1:e103.
22. Knoch D, Fehr E (2007) Resisting the power of temptations: The right prefrontal cortex and self-control. *Ann N Y Acad Sci* 1104:123–134.
23. Sakai K (2008) Task set and prefrontal cortex. *Annu Rev Neurosci* 31:219–245.
24. Kouneiher F, Charron S, Koechlin E (2009) Motivation and cognitive control in the human prefrontal cortex. *Nat Neurosci* 12:821–822.
25. McClure SM, et al. (2004) Neural correlates of behavioral preference for culturally familiar drinks. *Neuron* 44:379–387.
26. Li J, Xiao E, Houser D, Montague PR (2009) Neural responses to sanction threats in two-party economic exchange. *Proc Natl Acad Sci USA* 106:16835–16840.
27. Doll BB, Jacobs WJ, Sanfey AG, Frank MJ (2009) Instructional control of reinforcement learning: A behavioral and neurocomputational investigation. *Brain Res* 1299:74–94.
28. Biele G, Rieskamp J, Gonzalez R (2009) Computational models for the combination of advice and individual learning. *Cogn Sci* 33:206–242.
29. Delgado MR, Nystrom LE, Fissell C, Noll DC, Fiez JA (2000) Tracking the hemodynamic responses to reward and punishment in the striatum. *J Neurophysiol* 84:3072–3077.
30. de Quervain DJ, et al. (2004) The neural basis of altruistic punishment. *Science* 305: 1254–1258.
31. Kuhnen CM, Knutson B (2005) The neural basis of financial risk taking. *Neuron* 47: 763–770.
32. Delgado MR (2007) Reward-related responses in the human striatum. *Ann N Y Acad Sci* 1104:70–88.
33. Glimcher PW (2008) *Neuroeconomics: Decision Making and the Brain* (Academic Press, Burlington, MA), p 552.
34. Hare TA, Camerer CF, Rangel A (2009) Self-control in decision-making involves modulation of the vmPFC valuation system. *Science* 324:646–648.
35. Liu Z, Murray EA, Richmond BJ (2000) Learning motivational significance of visual cues for reward schedules requires rhinal cortex. *Nat Neurosci* 3:1307–1315.
36. Liu Z, Richmond BJ (2000) Response differences in monkey TE and perirhinal cortex: Stimulus association related to reward schedules. *J Neurophysiol* 83:1677–1692.
37. Mogami T, Tanaka K (2006) Reward association affects neuronal responses to visual stimuli in macaque te and perirhinal cortices. *J Neurosci* 26:6761–6770.
38. Daw ND, Niv Y, Dayan P (2005) Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat Neurosci* 8:1704–1711.
39. Dayan P, Daw ND (2008) Decision theory, reinforcement learning, and the brain. *Cogn Affect Behav Neurosci* 8:429–453.
40. McClure SM, Daw ND, Montague PR (2003) A computational substrate for incentive salience. *Trends Neurosci* 26:423–428.
41. O'Doherty JP, Dayan P, Friston K, Critchley H, Dolan RJ (2003) Temporal difference models and reward-related learning in the human brain. *Neuron* 38:329–337.
42. O'Doherty J, et al. (2004) Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science* 304:452–454.
43. Hampton AN, Bossaerts P, O'Doherty JP (2008) Neural correlates of mentalizing-related computations during strategic interactions in humans. *Proc Natl Acad Sci USA* 105:6741–6746.
44. O'Doherty J, Kringelbach ML, Rolls ET, Hornak J, Andrews C (2001) Abstract reward and punishment representations in the human orbitofrontal cortex. *Nat Neurosci* 4: 95–102.
45. O'Doherty J, Rolls ET, Francis S, Bowtell R, McGlone F (2001) Representation of pleasant and aversive taste in the human brain. *J Neurophysiol* 85:1315–1321.
46. Knutson B, Fong GW, Adams CM, Varner JL, Hommer D (2001) Dissociation of reward anticipation and outcome with event-related fMRI. *Neuroreport* 12:3683–3687.
47. Aharon I, et al. (2001) Beautiful faces have variable reward value: fMRI and behavioral evidence. *Neuron* 32:537–551.
48. Anderson AK, et al. (2003) Dissociated neural representations of intensity and valence in human olfaction. *Nat Neurosci* 6:196–202.
49. Small DM, et al. (2003) Dissociation of neural representation of intensity and affective valuation in human gustation. *Neuron* 39:701–711.
50. Zink CF, Pagnoni G, Martin-Skurski ME, Chappelow JC, Berns GS (2004) Human striatal responses to monetary reward depend on saliency. *Neuron* 42:509–517.
51. Spitzer M, Fischbacher U, Herrnberger B, Grön G, Fehr E (2007) The neural signature of social norm compliance. *Neuron* 56(1):185–196.
52. Delgado MR, Gillis MM, Phelps EA (2008) Regulating the expectation of reward via cognitive strategies. *Nat Neurosci* 11:880–881.
53. McClure SM, Laibson DI, Loewenstein G, Cohen JD (2004) Separate neural systems value immediate and delayed monetary rewards. *Science* 306:503–507.
54. Sanfey AG, Rilling JK, Aronson JA, Nystrom LE, Cohen JD (2003) The neural basis of economic decision-making in the Ultimatum Game. *Science* 300:1755–1758.
55. Gläscher J, Daw N, Dayan P, O'Doherty JP (2010) States versus rewards: Dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron* 66:585–595.