Neuron
Article

A Neural Signature of Hierarchical Reinforcement Learning

José J.F. Ribas-Fernandes,[1,2] Alec Solway,[1] Carlos Diuk,[1] Joseph T. McGuire,[3] Andrew G. Barto,[4] Yael Niv,[1,5] and Matthew M. Botvinick[1,5,*]
[1]Princeton Neuroscience Institute, Princeton University, Princeton, NJ 08540, USA
[2]Champalimaud Neuroscience Programme, Champalimaud Foundation, 1400-038 Lisbon, Portugal
[3]Department of Psychology, University of Pennsylvania, Philadelphia, PA 19104, USA
[4]Department of Computer Science, University of Massachusetts Amherst, Amherst, MA 01002, USA
[5]Department of Psychology, Princeton University, Princeton, NJ 08540, USA
*Correspondence: matthewb@princeton.edu
DOI 10.1016/j.neuron.2011.05.042

SUMMARY

Human behavior displays hierarchical structure: simple actions cohere into subtask sequences, which work together to accomplish overall task goals. Although the neural substrates of such hierarchy have been the target of increasing research, they remain poorly understood. We propose that the computations supporting hierarchical behavior may relate to those in hierarchical reinforcement learning (HRL), a machine-learning framework that extends reinforcement-learning mechanisms into hierarchical domains. To test this, we leveraged a distinctive prediction arising from HRL. In ordinary reinforcement learning, reward prediction errors are computed when there is an unanticipated change in the prospects for accomplishing overall task goals. HRL entails that prediction errors should also occur in relation to task *subgoals*. In three neuroimaging studies we observed neural responses consistent with such subgoal-related reward prediction errors, within structures previously implicated in reinforcement learning. The results reported support the relevance of HRL to the neural processes underlying hierarchical behavior.

INTRODUCTION

In recent years computational reinforcement learning (RL) (Sutton and Barto, 1998) has provided an indispensable framework for understanding the neural substrates of learning and decision making (Niv, 2009), shedding light on the functions of dopaminergic and striatal nuclei, among other structures (Barto, 1995; Montague et al., 1996; Schultz et al., 1997). However, to date, ideas from RL have been applied mainly in very simple task settings, leaving it unclear whether related principles might pertain in cases of more complex behavior (for a discussion, see Daw and Frank, 2009; Dayan and Niv, 2008). Hierarchically structured behavior provides a particularly interesting test case, not only because hierarchy plays an important role in human action (Cooper and Shallice, 2000; Lashley, 1951), but also because there exist RL algorithms specifically designed to operate in a hierarchical context (Barto and Mahadevan, 2003; Dietterich, 1998; Parr and Russell, 1998; Sutton et al., 1999).

Several researchers have proposed that such hierarchical reinforcement learning (HRL) algorithms may be relevant to understanding brain function, and a number of intriguing parallels to existing neuroscientific findings have been noted (Botvinick, 2008; Botvinick et al., 2009; Diuk et al., 2010, Soc. Neurosci., abstract, 907.14/KKK47 Badre and Frank, 2011; Haruno and Kawato, 2006). However, the relevance of HRL to neural function stands in need of empirical test.

In traditional RL (Sutton and Barto, 1998), the agent selects among a set of elemental actions, typically interpreted as relatively simple motor behaviors. The key innovation in HRL is to expand the set of available actions so that the agent may now opt to perform not only elemental actions, but also multiaction subroutines, containing sequences of lower-level actions, as illustrated in Figure 1 (for a fuller description, see Experimental Procedures and Botvinick et al., 2009).

Learning in HRL occurs at two levels. At a global level, the agent learns to select actions and subroutines so as to efficiently accomplish overall task goals. A fundamental assumption of RL is that goals are defined by their association with reward, and thus, the objective at this level is to discover behavior that maximizes long-term cumulative reward. Progress toward this objective is driven by temporal-difference (TD) procedures drawn directly from ordinary RL: following each action or subroutine, a reward prediction error (RPE) is generated, indicating whether the behavior yielded an outcome better or worse than initially predicted (see Figure 1 and Experimental Procedures), and this prediction error signal is used to update the behavioral policy. Importantly, outcomes of actions are evaluated with respect to the global goal of maximizing long-term reward.

At a second level, the problem is to learn the subroutines themselves. Intuitively, useful subroutines are designed to accomplish internally defined subgoals (Singh et al., 2005). For example, in the task of making coffee, one sensible subroutine would aim at adding cream. HRL makes the important assumption that the attainment of such subgoals is associated with a special form of reward, labeled *pseudo-reward* to distinguish it from "external" or primary reward. The distinction is critical because subgoals may not themselves be associated with primary reward. For example, adding cream to coffee may bring

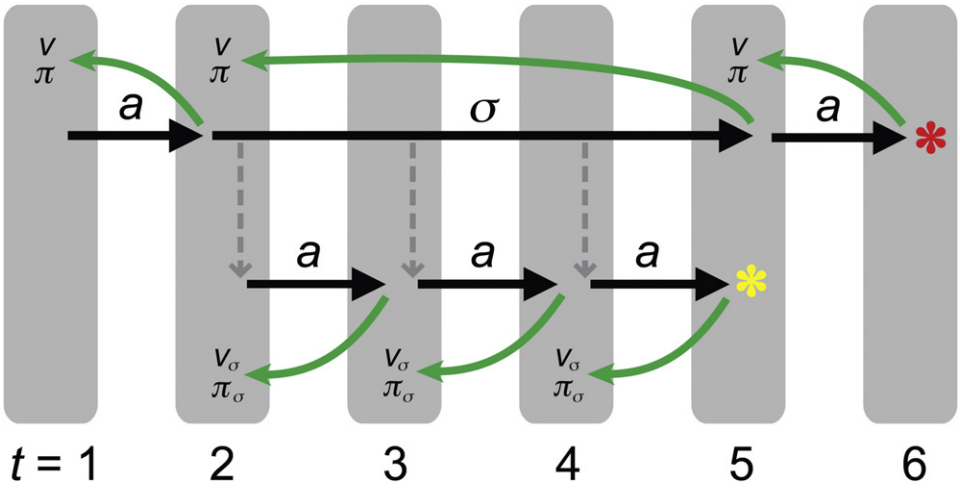

Figure 1. Illustration of HRL Dynamics

At *t*1, a primitive action (a) is selected. Based on the consequent state, an RPE is computed (green arrow from *t*2 to *t*1), and used to update the action policy (π) for the preceding state, as well as the value (V) of that state (an estimate of the expected future reward, when starting from that state). At *t*2 a subroutine (σ) is selected and remains active through t5. Until then, primitive actions are selected as dictated by σ (lower tier). A PPE is computed after each (lower green arrows from t5 to t2), and used to update the subroutine-specific action policy (π_σ) and state values (V_σ). These PPEs are computed with respect to pseudo-reward received at the end of the subroutine (yellow asterisk). Once the subgoal state of σ is reached, σ is terminated. An RPE is computed for the entire subroutine (upper green arrow from t5 to t2), and used to update the value and policy, V and π, associated with the state in which σ was initiated. A new action is then selected at the top level, yielding primary reward (red asterisk). Adapted from Botvinick et al. (2009).

one closer to that rewarding first sip, but is not itself immediately rewarding. In an HRL context, accomplishment of this subgoal would yield pseudo-reward, but not primary reward.

Once the HRL agent enters a subroutine, prediction error signals indicate the degree to which each action has carried the agent toward the currently relevant subgoal and its associated pseudo-reward (see Figure 1 and Experimental Procedures). Note that these subroutine-specific prediction errors are unique to HRL. In what follows, we refer to them as pseudo-reward prediction errors (PPEs), reserving "reward prediction error" for prediction errors relating to primary reward.

In order to make these points concrete, consider the video game illustrated in Figure 2, which is based on a benchmark task from the computational HRL literature (Dietterich, 1998). Only the colored elements in the figure appear in the task display. The overall objective of the game is to complete a "delivery" as quickly as possible, using joystick movements to guide the truck first to the package and from there to the house. It is self-evident how this task might be represented hierarchically, with delivery serving as the (externally rewarded) top-level goal and acquisition of the package as an obvious subgoal. For an HRL agent, delivery would be associated with primary reward and acquisition of the package with pseudo-reward. (This observation is not meant to suggest that the task *must* be represented hierarchically. Indeed, it is an established point in the HRL literature that any hierarchical policy has an equivalent nonhierarchical or flat policy, as long as the underlying decision problem satisfies the Markov property.) Our neuroimaging experiments proceeded on the assumption that participants would represent the delivery task hierarchically. However, as we discuss later, the neuroimaging results themselves, together with results from a behavioral experiment, provided convergent evidence for the validity of this assumption. See Supplemental Experimental Procedures, available online, for further discussion.

Consider now a version of the task in which the package sometimes unexpectedly jumps to a new location before the truck reaches it. According to RL, a jump to point A in the figure, or any location within the ellipse shown, should trigger a positive RPE because the total distance that must be covered in order to deliver the package has decreased. (Note that we assume temporal discounting, which implies that attaining the goal faster is more rewarding. We also assume that current subgoal and goal distances are always immediately known, as they were for our experimental participants from the task display.) By the same token, a jump to point B or any other exterior point should trigger a negative RPE. Cases C, D, and E are quite different. Here, there is no change in the overall distance to the goal, and so no RPE should be triggered, either in standard RL or in HRL. However, in case C the distance to the subgoal has decreased. Thus, according to HRL, a jump to this location should trigger a positive PPE. Similarly, a jump to location D should trigger a negative PPE (note that location E is special, being the only location that should trigger neither an RPE nor a PPE). These points are illustrated in Figure 2 (right), which shows RPE and PPE time courses from simulations of the delivery task based on standard RL and HRL (for simulation methods, see Experimental Procedures).

These points translate directly into neuroscientific predictions. Previous research has revealed neural correlates of the RPE in numerous structures (Breiter et al., 2001; Hare et al., 2008; Holroyd and Coles, 2002; Holroyd et al., 2003; O'Doherty et al., 2003; Ullsperger and von Cramon, 2003; Yacubian et al., 2006). HRL predicts that neural correlates should also exist for the PPE. To test this, we had neurologically normal participants

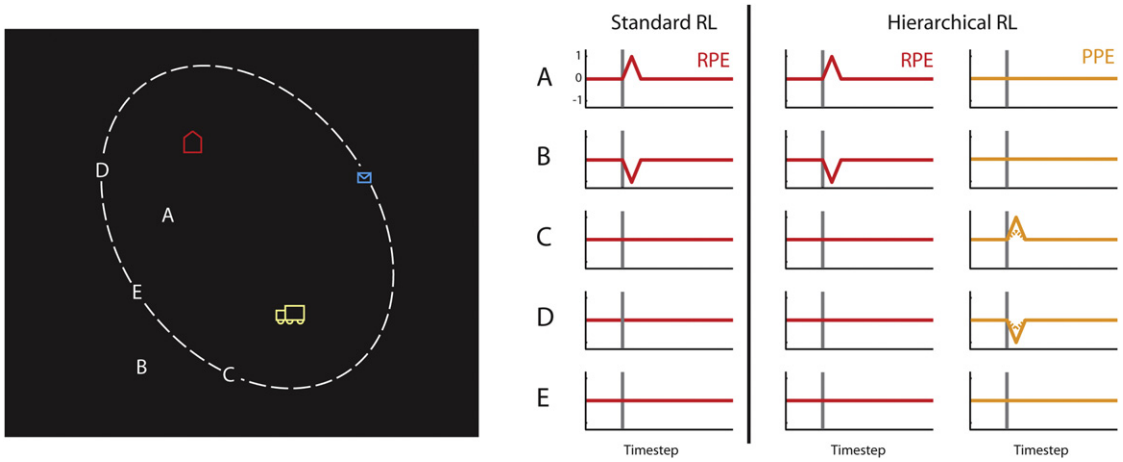

Figure 2. Task and Predictions from HRL and RL

Left view is task display and underlying geometry of the delivery task. Right view shows prediction-error signals generated by standard RL and by HRL in each category of jump event. Gray bars mark the time step immediately preceding a jump event. Dashed time courses indicate the PPE generated in C and D jumps that change the subgoal's distance by a smaller amount. For simulation methods, see Experimental Procedures.

perform the delivery task from Figure 2 while undergoing EEG and, in two further experiments, fMRI.

RESULTS

EEG Experiment

The EEG experiment included 9 participants, who performed the delivery task for a total of 60 min (190 delivery trials on average per participant). One-third of trials involved a jump event of type D from Figure 2; these events were intended to elicit a negative PPE. Earlier EEG research indicates that ordinary negative RPEs trigger a midline negativity typically centered on lead Cz, sometimes referred to as the feedback-related negativity or FRN (Holroyd and Coles, 2002; Holroyd et al., 2003; Miltner et al., 1997). Based on HRL, we predicted that a similar negativity would occur following the critical jumps (type D) in our task. To provide a baseline for comparison, another third of the trials involved jump events of type E.

Stimulus-aligned EEG averages indicated that class D-jump events triggered a phasic negativity in the EEG ($p < 0.01$ at Cz; Figure 3, left), relative to the E-jump control condition. (Like the ERP obtained in this study, the FRN sometimes takes the form of a relative negativity occupying the positive voltage domain, rather than absolute negativity. For germane examples, see Nieuwenhuis et al., 2005; Yeung et al., 2005.) Like the FRN, this negativity was largest in the fronto-central midline leads (including Cz, see Figure 3, right), and although the observed negativity peaked later than the typical FRN, its timing is consistent with studies of equivalent complexity of feedback (Baker and Holroyd, 2011).

fMRI Experiments

In our first fMRI experiment, a group of 30 new participants performed a slightly different version of the delivery task, again designed to elicit negative PPEs. As in the EEG experiment, one-third of trials included a jump of type D (as in Figure 2),

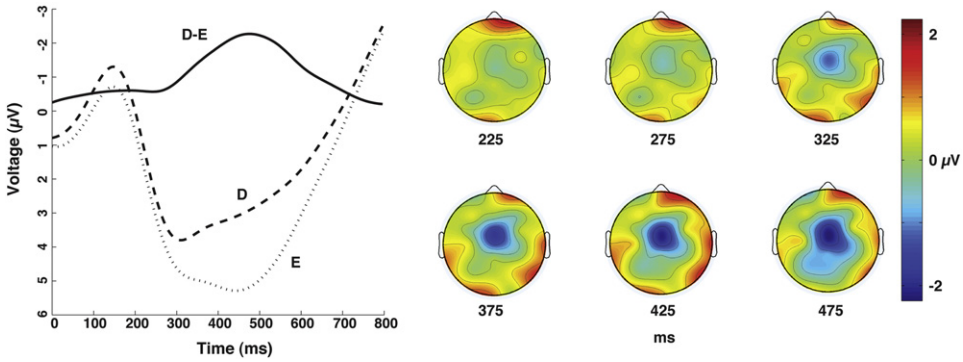

Figure 3. Results of EEG Experiment

Left view shows evoked potentials at electrode Cz, aligned to jump events, averaged across participants. D and E refer to jump destinations in Figure 2. The data series labeled D-E shows the difference between curves D and E, isolating the PPE effect. Right view is scalp topography for condition D, with baseline condition E subtracted (topography plotted on the same grid used in Yeung et al. [2005]).

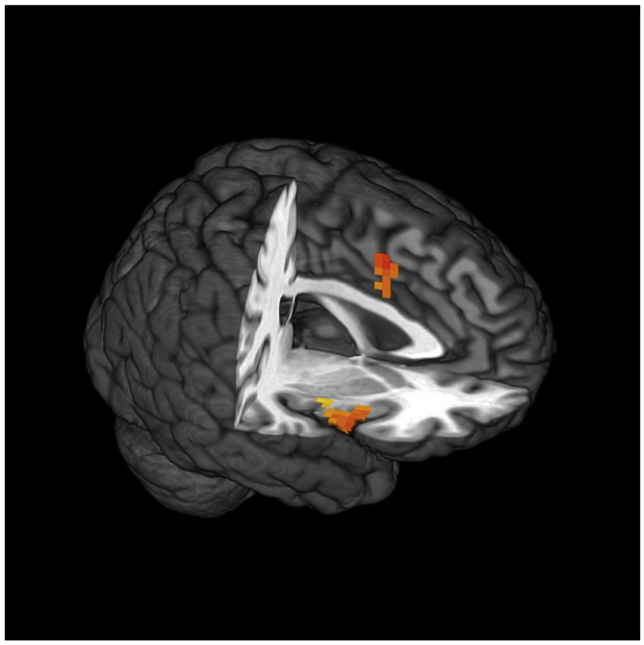

Figure 4. Results of fMRI Experiment 1

Shown are regions displaying a positive correlation with the PPE, independent of subgoal displacement. Talairach coordinates of peak are 0, 9, and 39 for the dorsal ACC, and 45, 12, and 0 for right anterior insula. Not shown are foci in left anterior insula (−45, 9, −3) and lingual gyrus (0, −66, 0). Color indicates general linear model parameter estimates, ranging from 3.0×10^{-4} (palest yellow) to 1.2×10^{-3} (darkest orange).

and another third included a jump of type E. Type D jumps, by increasing the distance to the subgoal, were again intended to trigger a PPE. However, in the fMRI version of the task, unlike the EEG version, the exact increase in subgoal distance varied across trials. Therefore, type D jumps were intended to induce PPEs that varied in magnitude (Figure 2). Our analyses took a model-based approach (O'Doherty et al., 2007), testing for regions that showed phasic activation correlating positively with predicted PPE size.

A whole-brain general linear model analysis, thresholded at $p < 0.01$ (cluster-size thresholded to correct for multiple comparisons), revealed such a correlation in the dorsal anterior cingulate cortex (ACC; Figure 4). This region has been proposed to contain the generator of the FRN (Holroyd and Coles, 2002, although see Nieuwenhuis et al., 2005 and Discussion below). In this regard the fMRI result is consistent with the result of our EEG experiment. The same parametric fMRI effect was also observed bilaterally in the anterior insula, a region often coactivated with the ACC in the setting of unanticipated negative events (Phan et al., 2004). The effect was also detected in right supramarginal gyrus, the medial part of lingual gyrus, and, with a negative coefficient, in the left inferior frontal gyrus. However, in a follow-up analysis we controlled for subgoal displacement (e.g., the distance between the original package location and point D in Figure 2), a nuisance variable moderately correlated, across trials, with the change in distance to subgoal. Within this analysis only the ACC ($p < 0.01$), bilateral anterior insula ($p < 0.01$ left, $p < 0.05$ right), and right lingual gyrus ($p < 0.01$) continued to show significant correlations with the PPE.

In a series of region-of-interest (ROI) analyses, we focused in on additional neural structures that, like the ACC, have been previously proposed to encode negative RPEs: the habenular complex (Salas et al., 2010; Ullsperger and von Cramon, 2003), nucleus accumbens (NAcc) (Seymour et al., 2007), and amygdala (Breiter et al., 2001; Yacubian et al., 2006). (These analyses were intended to bring greater statistical power to bear on these regions, in part because their small size may have undermined our ability to detect activation in them in our whole-brain analysis, where a cluster-size threshold was employed.) The habenular complex was found to display greater activity following type D than type E jumps ($p < 0.05$), consistent with the idea that this structure is also engaged by negative PPEs. A comparable effect was also observed in the right, though not left, amygdala ($p < 0.05$).

In the NAcc, where some studies have observed deactivation accompanying negative RPEs (Knutson et al., 2005), no significant PPE effect was observed. However, it should be noted that NAcc deactivation with negative RPEs has been an inconsistent finding in previous work (for example, see Cooper and Knutson, 2008; O'Doherty et al., 2006). More robust is the association between NAcc activation and positive RPEs (Hare et al., 2008; Niv, 2009; Seymour et al., 2004). To test this directly, we ran a second, smaller fMRI study designed to elicit positive PPEs, specifically looking for activation within a NAcc ROI. A total of 14 participants performed the delivery task, with jumps of type C (in Figure 2) occurring on one-third of trials and jumps of type E on another third. As described earlier, a positive PPE is predicted to occur in association with type C jumps, and in this setting significant activation ($p < 0.05$) was observed in the right (though not left) NAcc, scaling with predicted PPE magnitude.

Behavioral Experiment

We have characterized the results from our EEG and fMRI experiments as displaying a "signature" of HRL, in the sense that the PPE signal is predicted by HRL but not by standard RL algorithms (Figure 2). However, there is an important caveat that we now consider. In our neuroimaging experiments we assumed that reaching the goal (the house) would be associated with primary reward. (The same points hold if "primary reward" is replaced with "secondary" or "conditioned reinforcement.") We also assumed that reaching the subgoal (the package) was not associated with primary reward but only with pseudoreward. However, what if participants did attach primary reward to the subgoal? If this were the case, it would present a difficulty for the interpretation of our neuroimaging results because it would lead standard RL to predict an RPE in association with events that change only subgoal distance (including C and D jumps in our neuroimaging task).

In view of these points, it was necessary to establish whether participants performing the delivery task did or did not attach primary reward to subgoal attainment. In order to evaluate this, we devised a modified version of the task. Here, 22 participants delivered packages as before, though without jump events. However, at the beginning of each delivery trial, two packages were presented in the display, which defined paths that could

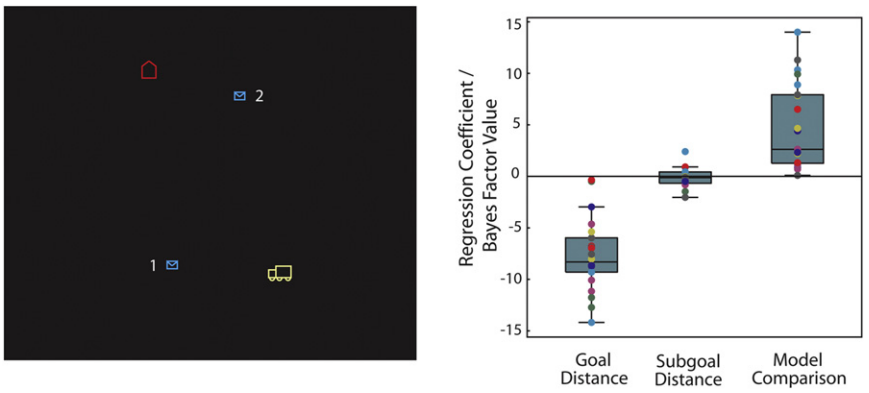

Figure 5. Results of Behavioral Experiment
Left view is an example of a choice display. Subgoal 1 would always be on an ellipse defined by the house and the truck. In this example subgoal 2 has smaller overall distance to the goal and larger distance to the truck relative to subgoal 1 (labels not shown to participants). Right view shows results of logistic regression on choices and of the comparison between two RL models. Choices were driven significantly by the ratio of distances of the goal of the two subgoals (left box, central mark is the median, edges correspond to 25th and 75th percentiles, whiskers to extreme values, outliers to individual dots outside box and whiskers; each colored dot represents a single participant's data), whereas the ratio of distances to subgoal did not significantly explain participant's choices (middle box). Bayes factors favored the model with only reward for goal attainment and no reward for subgoal against the one with reward for subgoal and goal attainment (right box).

differ both in terms of their subgoal distance and the overall distance to the goal (Figure 5, left). Participants indicated with a key press which package they preferred to deliver.

We reasoned that if goal attainment were associated with primary reward, then (assuming ordinary temporal discounting) the overall goal distance associated with each of the two packages should influence choice. More importantly, if we were correct in our assumption that subgoal attainment carried no primary reward, then choice should not be influenced by subgoal distance, i.e., the distance from the truck to each of the two packages.

Participants' choices strongly supported both of these predictions. Logistic regression analyses indicated that goal distance had a strong influence on package choice ($M = -7.6$, $p < 0.001$; Figure 5, right; larger negative coefficients indicate a larger penalty on distances). However, subgoal distance exerted no appreciable influence on choice ($p = 0.43$), and the average regression coefficient was near zero (-0.16). The latter observation held even in a subset of trials where the two delivery options were closely matched in terms of overall distance (with ratios of overall goal distance between 0.8 and 1.2).

These behavioral results strongly favor our HRL account of delivery task, over a standard RL account. (The behavioral data are consistent with a standard RL model that attaches no reward to subgoal attainment, but as noted earlier, such a model offers no explanation for our neuroimaging results.) To further establish the point, we fit two computational models to individual subjects' choice data: (1) an HRL model, and (2) a standard RL model in which primary reward was attached to the subgoal (see Experimental Procedures). The mean Bayes factor across subjects—with values greater than one favoring the HRL model—was 4.31, and values across subjects differed significantly from one (two-tailed t test, $p < 0.001$; see Figure 5, right).

DISCUSSION

We predicted, based on HRL, that neural structures previously proposed to encode TD RPEs should also respond to PPEs—prediction errors tied to behavioral subgoals. Across three experiments using a task designed to elicit PPEs, without eliciting RPEs, we observed evidence consistent with this prediction. Negative PPEs were found to engage three structures previously reported to show activation with negative RPEs: ACC, habenula, and amygdala; and activation scaling with positive PPEs was observed in right NAcc, a location frequently reported to be engaged by positive RPEs.

Of course the association of these neural responses with the relevant task events does not uniquely support an interpretation in terms of HRL (see Poldrack, 2006). However, aspects of either the task or the experimental results do militate against the most tempting alternative interpretations. Our behavioral study provided evidence against primary reward at subgoal attainment, closing off an interpretation of the neuroimaging data in terms of standard RL. Given previous findings pertaining to the ACC, the effect we observed in this structure might be conjectured to reflect response conflict or error detection (Botvinick et al., 1999; Krigolson and Holroyd, 2006; Yeung et al., 2004). However, additional analyses of the EEG data (see Figure S2 and Supplemental Experimental Procedures) indicated that the PPE effect persisted even after controlling for response accuracy and for response latency, each commonly regarded as an index of response conflict.

Another alternative that must be addressed relates to spatial attention. Jump events in our neuroimaging experiments presumably triggered shifts in attention, often complete with eye movements, and it is important to consider the possibility that differences between conditions on this level may have contributed to our central findings. Although further experiments may be useful in pinning down the precise role of attention in our task, there are several aspects of the present results that argue against an interpretation based purely on attention. Note that, in previous EEG research, exogenous shifts of attention have been associated with a midline positivity, the amplitude of which grows with stimulus eccentricity (Yamaguchi et al., 1995). (A midline negativity has been reported in at least one study focusing on endogenous attention (Grent-'t-Jong and Woldorff

[2007]), but the timing of this potential differed dramatically from the difference wave in our EEG study, peaking at 1000–1200 ms poststimulus, hundreds of milliseconds after our effect ended.) In fact we observed such a positivity in our own data, in Cz, when we compared jump events (D and E) against occasions where the subgoal stayed put, an analysis specifically designed to uncover attentional effects (Figure S3). In contrast the PPE effect in our data took the form of a negative difference wave (Figure 3), consistent with the predictions of HRL and contrary to those proceeding from previous research on attention.

Our fMRI results also resist an interpretation based on spatial attention alone. As detailed in the Supplemental Experimental Procedures, we did find activation in or near the frontal eye fields and in the superior parietal cortex—regions classically associated with shifts of attention (Corbetta et al., 2008)—in an analysis contrasting all jump events with trials where the subgoal remained in its original location (Figure S4). However, as reported above, activity in these regions did not show any significant correlation with our PPE regressor (Figure 4).

If one does adopt an HRL-based interpretation of the present results, then several interesting questions follow. Given the prevailing view that TD RPEs are signaled by phasic changes in dopaminergic activity (Schultz et al., 1997), one obvious question is whether the PPE might be signaled via the same channel. ACC activity in association with negative RPEs has been proposed to reflect phasic reductions in dopaminergic input (Holroyd and Coles, 2002), and the habenula has been proposed to provide suppressive input to midbrain dopaminergic nuclei (Christoph et al., 1986; Matsumoto and Hikosaka, 2007). Thus, the implication of the ACC and habenula in the present study, as well as the involvement of the NAcc—another structure that has been proposed to show activity related to dopaminergic input (Nicola et al., 2000)—provides tentative, indirect support for dopaminergic involvement in HRL. At the same time, it should be noted that some ambiguity surrounds the role of dopamine in driving reward-outcome responses, particularly within the ACC (for a detailed review, see Jocham and Ullsperger, 2009). Indeed, some disagreement still exists concerning whether the dorsal ACC is responsible for generating the FRN (compare Holroyd et al., 2004; Nieuwenhuis et al., 2005; van Veen et al., 2004). Thus, the present findings must be interpreted with appropriate circumspection. Above all, it should be noted that our HRL-based interpretation does not necessarily require a role for dopamine in generating the observed neural events. Indeed, if the PPE were conveyed via phasic dopaminergic signaling, this would give rise to an interesting computational problem because proper credit assignment would require discrimination between PPE and RPE signals (for discussion, see Botvinick et al., 2009).

Another important question for further research concerns the relation between the present findings and recent data concerning the representation of action hierarchies in the dorsolateral prefrontal cortex (Badre, 2008; Botvinick, 2008). Neuroimaging and neuropsychological studies have lately given rise to the idea that the prefrontal cortex may display a rostrocaudal functional topography, which separates out task representations based on some measure of abstractness (Badre et al., 2009; Christoff et al., 2009; Grafman, 2002; Kouneiher et al., 2009). One speculation, which could be tested through further research, is that HRL-like mechanisms might be responsible for shaping such representations and gating them into working memory in an adaptive fashion (see Botvinick et al., 2009; Reynolds and O'Reilly, 2009).

One final challenge for future research is to test predictions from HRL in settings involving learning-driven changes in action selection. As in many neuroscientific studies focusing on RL mechanisms, our task looked at prediction errors in a setting where behavioral policies were more or less stable. It may also prove useful to study the dynamics of learning in hierarchically structured tasks, as a further test of the relevance of HRL to neural function (see Diuk et al., 2010, Soc. Neurosci., abstract, 907.14/KKK47; Badre and Frank, 2011).

EXPERIMENTAL PROCEDURES

An HRL Model of the Delivery Task

To make our computational predictions explicit, we implemented both a standard and a hierarchical RL model of the delivery task, based on the approach laid out in Botvinick et al. (2009). Simulations were performed in MATLAB (The MathWorks, Natick, MA); the relevant code is available for download from http://www.princeton.edu/~matthewb.

For the standard RL agent, the state on each step t, labeled s_t, was represented by the goal distance (*gd*), the distance from the truck to the house, via the package, in units of navigation steps. For the HRL agent the state was represented by two numbers: *gd* and the subgoal distance (*sd*), i.e., the distance between the truck and the package. Goal attainment yielded a reward (r) of one for both agents, and subgoal attainment a pseudo-reward (p) of one for the HRL agent. On each step of the task, the agent was assumed to act optimally, i.e., to take a single step directly toward the package or, later in the task, toward the house. The HRL agent was assumed to select a subroutine (σ) for attaining the package, which also resulted in direct steps toward this subgoal (for details of subtask specification and selection, see Figure 1 and Botvinick et al., 2009; Sutton et al., 1999).

For the standard RL agent, the state value at time t, $V(t)$, was defined as γ^{gd}, using a discount factor $\gamma = 0.9$. Thus, the RPE on steps prior to goal attainment was:

$$RPE = r_{t+1} + \gamma V(s_{t+1}) - V(s_t) = \gamma^{1+gd_{t+1}} - \gamma^{gd_t}. \quad (1)$$

The HRL agent calculated RPEs in the same manner but also calculated PPEs during execution of the subroutine σ. These were based on a subroutine-specific value function (see Botvinick et al., 2009; Sutton et al., 1999), defined as $V_\sigma(s_t) = \gamma^{sd_t}$.

Thus, the PPE on each step prior to subgoal attainment was:

$$PPE = \rho_{t+1} + \gamma V_\sigma(s_{t+1}) - V_\sigma(s_t) = \gamma^{1+sd_{t+1}} - \gamma^{sd_t}. \quad (2)$$

To generate the data shown in Figure 2, we imposed initial distances (*gd*, *sd*) equaling 949 and 524. Following two task steps in the direction of the package, at a point with distances 849 and 424, in order to represent jump events distances were changed to 599 and 424 for jump type A, 1449 and 424 for type B, 849 and 124 for type C, 849 and 724 for type D, and 849 and 424 for type E. Dashed data series in Figure 2 were generated with jumps to 849 and 236 for type C and 849 and 574 for type D.

EEG Experiment

Participants

All experimental procedures were approved by the Institutional Review Board of Princeton University. Participants were recruited from the university community, and all gave their informed consent. Nine participants were recruited (ages 18–22 years, M = 19.7, 4 males, all right handed). All received course credit as compensation, and in addition received a monetary bonus based on their performance in the task.

Task and Procedure

Participants sat at a comfortable distance from a shielded CRT display in a dimly lit, sound-attenuating, electrically shielded room. A joystick was held in the right hand (Logitech International, Romanel-sur-Morges, Switzerland).

The computerized task was coded using MATLAB (The MathWorks) and the MATLAB Psychophysics Toolbox, version 3 (Brainard, 1997). On each trial, three display elements appeared: a truck, a package, and a house (Figure S1A). These objects occupied the vertices of a virtual triangle with vertices at pixel coordinates 0 and 180, 150 and 30, and 0 and 180, relative to the center of the screen (resolution 1024 × 768) but assuming a random new rotation and reflection at the onset of each trial. The task was to move the truck first to the package and then to the house. Each joystick movement displaced the truck a fixed distance of 50 pixels. For reasons given below the orientation of the truck was randomly chosen after every such translation, and participants were required to tailor their joystick responses to the truck's orientation, as if they were facing its steering wheel (Figure S1A). For example if the front of the truck were oriented toward the bottom of the screen, rightward movement of the joystick would move the truck to the left. This aspect of the task was intended to ensure that intensive spatial processing occurred at each step of the task, rather than only following subgoal displacements.

Responses were registered when the joystick was tilted beyond half its maximum displacement (Figure S1A). Between responses the participant was required to restore the joystick to a central position (Figures S1A and S1B). When the truck passed within 30 pixels of the package, the package moved inside the truck icon and remained there for subsequent moves. When the truck containing the package passed within 35 pixels of the house, the display cleared, and a message reading "10¢" appeared for a duration of 300 ms (participants were paid their cumulative earnings at the end of the experiment). A central fixation cross then appeared for 700 ms before the onset of the next trial.

On every trial, after the first, second, or third truck movement, a brief tone occurred, and the package flashed for an interval of 200 ms, during which any joystick inputs were ignored. On one-third of such occasions, the package remained in its original location. On the remaining trials, at the onset of the tone, the package jumped to a new location. In half of such cases, the distance between the package's new position and the truck position was unchanged by the jump (case E in Figure 2 of the main text). In the remaining cases the distance from the truck to the package was increased by the jump, although the total distance from the truck to the house (via the package) remained the same (case D in Figure 2). In these cases the jump always carried the package across an imaginary line connecting the truck and the house, and always resulted in a package-to-house distance of 160 pixels. In all three conditions the package would be on an ellipse defined by the locations of the old subgoal, the house, and the position of the truck at the time of the jump. By the definition of an ellipse, overall distance to the house was preserved.

At the outset of the experiment, each participant completed a 15 min training session, which was followed by the hour-long EEG testing session. Participants completed 190 trials on average (range 128–231). Trials were grouped into blocks, each containing six trials: two trials in which the position of the package did not change, two involving type E jumps, and two type D jumps. The order in which trials of a particular type occurred was pseudorandom within a block. Participants were given an opportunity to rest for a brief period between task blocks.

Data Acquisition

EEG data were recorded using Neuroscan (Charlotte, NC) caps with 128 electrodes and a Sensorium (Charlotte, VT) EPA-6 amplifier. The signal was sampled at 1000 Hz. All data were referenced online to a chin electrode, and after excluding bad channels were rereferenced to the average signal across all remaining channels (Hestvik et al., 2007). EOG data were recorded using a single electrode placed below the left eye. Ocular artifacts were detected by thresholding a slow-moving average of the activity in this channel, and trials with artifacts were not included in the analysis. Less than four trials per subject matched this criterion and were excluded from the analysis (less than two per condition).

Data Analysis

Epochs of 1000 ms (200 ms baseline) were extracted from each trial, time locked to the package's change in position. The mean level of activity during the baseline interval was subtracted from each epoch. Trials containing type D jump were separated from trials containing jumps of type E, and ERPs were computed for each condition and participant by averaging the corresponding epochs. The ERPs shown in Figure 3 (main text) were computed by averaging across participants. The PPE effect was quantified in electrode Cz (following Holroyd and Coles, 2002).

The PPE effect was quantified for each subject by taking the mean voltage during the time window from 200 to 600 ms following each jump, for the two jump types. A one-tailed paired t test was used to evaluate the hypothesis that type D jumps elicited a more negative potential than type E jumps. For comparability with previous studies, topographic plots are shown for electrodes FP1, FP2, AFz, F3, Fz, F4, FT7, FC3, FCz, FC4, FT8, T7, C3, Cz, C4, T8, TP7, CP3, CPz, CP4, TP8, P7, P3, Pz, P4, P8, O1, Oz, and O2 (as in Yeung et al. [2005], F7 and F8 were an exception, given that the used cap did not have these electrode locations).

fMRI Experiments

Participants

Participants were recruited from the university community and all gave their informed consent. For the first fMRI experiment, 33 participants were recruited (ages 18–37 years, M = 21.2, 20 males, all right handed). Three participants were excluded: two because of technical problems and one who was unable to complete the task in the available time. For the second experiment, 15 participants were recruited (ages 18–25 years, M = 20.5, 11 males, all were right handed). One participant was excluded for failure to complete the task in the available time. All participants received monetary compensation at a departmental standard rate. Participants in the second experiment also received a small monetary bonus based on task performance.

Task and Procedure

An MR-compatible joystick (MagConcept, Redwood City, CA) was used. The task was identical to the one used in the EEG experiment, with the following exceptions. For the first experiment initial positions of the icons were randomly assigned to the screen respecting a minimal distance of 150 pixels between icons. For the second experiment initial positions of the icons were rotations or reflections, varied randomly, of a preestablished arrangement of icons of a predetermined triangle with vertices truck (0, 200), package (151, −165), and house (0, −200) (coordinates are in pixels, referenced to the center of the screen). On type D jumps, the destination of the package was chosen randomly from all locations satisfying the conditions that they (1) increase truck-to-package distance, but (2) leave total path length to the goal (house) unchanged. The forced delay involved in the task interruption (tone, package flashing) totaled 900 ms. At the completion of each delivery, the message "Congratulations!" was displayed for 1000 ms (Figure S1D), followed by a fixation cross that remained on screen for 6000 ms.

The first fMRI experiment consisted of three parts: a 15 min behavioral practice outside the scanner, an 8 min practice inside the scanner during structural scan acquisition, and a third phase of approximately 45 min, where functional data were collected. During functional scanning, 90 trials were completed, in 6 runs of 15 trials each. At the beginning and end of each run, a central fixation cross was displayed for 10,000 ms. The average run length was 7.5 min (range 5.7–11).

The task and procedure in the second fMRI experiment were identical to those in the first, with the following exceptions. Type D jumps were replaced with type C jumps (see Figure 2 in the main text). In these cases, the distance between truck and package always decreased to 120 pixels. The message "10¢" appeared for 500 ms, indicating the bonus earned for that trial. Immediately following this, a fixation cross appeared for 2500 ms, followed by onset of the next trial. The average run length was 6.8 min (range 4.7–10.7).

Image Acquisition

Image acquisition protocols were the same for both experiments. Data were acquired with a 3 T Siemens Allegra (Malvern, PA) head-only MRI scanner, with a circularly polarized head volume coil. High-resolution (1 mm^3 voxels) T1-weighted structural images were acquired with an MP-RAGE pulse sequence at the beginning of the scanning session. Functional data were acquired using a high-resolution echo-planar imaging pulse sequence (3 × 3 × 3 mm voxels, 34 contiguous slices, 3 mm thick, interleaved acquisition, TR of 2000 ms, TE of 30 ms, flip angle 90°, field of view 192 mm, aligned

with the anterior commissure-posterior commissure plane). The first five volumes of each run were ignored.

Data Analysis

Data analysis was similar for both experiments. Data were analyzed using AFNI software (Cox, 1996). The T1-weighted anatomical images were aligned to the functional data. Functional data were corrected for interleaved acquisition using Fourier interpolation. Head motion parameters were estimated and corrected allowing six-parameter rigid body transformations, referenced to the initial image of the first functional run. A whole-brain mask for each participant was created using the union of a mask for the first and last functional images. Spikes in the data were removed and replaced with an interpolated data point. Data were spatially smoothed until spatial autocorrelation was approximated by a 6 mm FHWM Gaussian kernel. Each voxel's signal was converted to percent change by normalizing it based on intensity. The mean image for each volume was calculated and used later as baseline regressor in the general linear model, except in the ROI analysis where the mean image of the whole brain was not subtracted from the data. Anatomical images were used to estimate normalization parameters to a template in Talairach space (Talairach and Tournoux, 1988), using SPM5 (http://www.fil.ion.ucl.ac.uk/spm/). These transformations were applied to parameter estimates from the general linear model.

General Linear Model Analysis

For each participant we created a design matrix modeling experimental events and including events of no interest. At the time of an experimental event, we defined an impulse and convolved it with a hemodynamic response. The following regressors were included in the model: (a) an indicator variable marking the occurrence of all auditory tone/package flash events; (b) an indicator variable marking the occurrence of all jump events (spanning jump types E and D in Experiment 1 and types E and C in Experiment 2); (c) an indicator variable marking the occurrence of type D jumps (C jumps in Experiment 2); (d) a parametric regressor indicating the change in distance to subgoal induced by each D (or C) jumps, mean centered; (e and f) indicator variables marking subgoal and goal attainment; and (g) an indicator variable marking all periods of task performance, from the initial presentation of the icons to the end of the trial. Also included were head motion parameters, and first- to third-order polynomial regressors to regress out scanner drift effects. In Experiment 1, a global signal regressor was also included (comparable analyses omitting the global signal regressor yielded statistically significant PPE effects in the ACC, bilateral insula, and lingual gyrus, in locations highly overlapping with those reported in the main text).

Group Analysis (Experiment 1)

For each regressor and for each voxel, we tested the sample of 30 subject-specific coefficients against zero in a two-tailed t test. We defined a threshold of $p = 0.01$ and applied correction for multiple comparison based on cluster size, using Monte Carlo simulations as implemented in AFNI's AlphaSim. We report results at a corrected $p < 0.01$.

Follow-up Analysis (Experiment 1)

Our experimental prediction related to the change in distance between truck and package induced by type D-jump events, i.e., the change in distance to subgoal, or PPE effect. However, jump events also varied in the degree to which they displaced the package (i.e., the distance from its original position to its post-jump position), and this distance correlated moderately with the increase in subgoal distance. Therefore, it was necessary to evaluate whether the regions of activation identified in our primary GLM analysis might simply be responding to subgoal displacement (and possible attendant visuospatial or motor processes), rather than the increase in distance to subgoal. To this end, we looked at each area identified in the primary GLM, asking whether the area continued to a show significant PPE effect even after this regressor was made orthogonal to subgoal displacement. In order to avoid bias in this procedure, we employed a leave-one-out cross-validation approach, as follows. For every subgroup of 29 participants (from the total sample of 30), we reran the original GLM, identifying voxels that: (1) showed the PPE effect at significance threshold of $p = 0.05$ (cluster-size thresholded to compensate for multiple comparisons); and (2) fell within 33 mm of the peak-activation coordinates for one of the six clusters identified in our primary GLM (dorsal anterior cingulate, bilateral anterior insulae, left lingual gyrus, left inferior frontal gyrus, and right supramarginal gyrus). The resulting clusters were used as ROIs for the critical test. Focusing on the one subject omitted from each 29 subject subsample, we calculated the mean coefficient within each ROI for the PPE effect, after orthogonalizing the PPE regressor to subgoal displacement (and including subgoal displacement in the GLM). This yielded 30 coefficients per ROI. Each set was tested for difference from zero, using a two-tailed t test.

ROI Analysis

For the first fMRI experiment, we defined NAcc based on anatomical boundaries on a high-resolution T1-weighted image for each participant; habenula, using peak Talairach coordinates (5, 25, 8), guided by Ullsperger and von Cramon (2003), surrounded by a sphere with a radius of 6 mm (Salas et al., 2010); and amygdala, drawn using the Talairach atlas in AFNI. For the second experiment we defined NAcc in the same way as for the first experiment. Mean coefficients were extracted from these regions for each participant. Reported coefficients for all ROIs are from general linear model analyses without subtraction of global signal. The sample of 30 (or 14 for the second experiment) subject-specific coefficients was tested against zero in a two-tailed t test, with a threshold of $p < 0.05$.

Behavioral Experiment

Participants

A total of 22 participants were recruited from the Princeton University community (ages 18–22 years, 11 male). All provided informed consent and received a nominal payment.

Task and Procedure

The experiment was composed of three phases. In the first phase, participants completed ten deliveries, with the procedure matching that used in our fMRI studies. However, no jump events occurred in this or later phases of the experiment. The second phase consisted of ten further delivery trials. However, here, at the onset of each trial, the participant was required to choose between two packages (Figure 5). The location of the truck and the house was chosen randomly. The location of one package, designated subgoal one, was randomly positioned along an ellipse with the truck and house as its foci and a major-to-minor axis ratio of 5/3. The position of the other package, subgoal two, was randomly chosen, subject to the constraint that it fall at least 100 pixels from each of the other icons.

At the onset of each trial, each package would be highlighted with a change of color, twice (in alternation with the other package), for a period of 1.5 s. Highlighting order was counterbalanced across trials. During this period the participant was required to press a key to indicate his or her preferred package when that package was highlighted. After the key press, the chosen subgoal would change to a new color. At the end of the choice period, the unchosen subgoal was removed, and participants were expected to initiate the delivery task. The remainder of each trial proceeded as in phase one.

The third and main phase of the experiment included 100 trials. One-third of these, interleaved in random order with the rest, followed the profile of phase two trials. The remaining trials began as in phase two but terminated immediately following the package-choice period.

Data Analysis

To determine the influence of goal and subgoal distance on package choice, we conducted a logistic regression on the choice data from phase three. Regressors included (1) the ratio of the distances from the truck to subgoal one and subgoal two, and (2) the ratio of the distances from the truck to the house through subgoal one and subgoal two. To test for significance across subjects, we carried out a two-tailed t test on the population of regression coefficients.

To further characterize the results, we fitted two RL models to each participant's phase-three choice data. One model assigned primary reward only to goal attainment and so was indifferent to subgoal distance per se. A second model assigned primary reward to the subgoal as well to the goal.

Value in the first case was a discounted number of steps to the goal, and in the second case it was a sum of discounted number of steps to the subgoal and to the goal. Choice was modeled using a softmax function, including a free inverse temperature parameter. The fmincon function in MATLAB was employed to fit discount factor and inverse temperature parameters for both models and reward magnitude for subgoal attainment for the second model. We then compared the fits of the two models calculating Bayes factor for each participant and performing a two-tailed t test on the factors.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures and four figures and can be found with this article online at doi:10.1016/j.neuron.2011.05.042.

ACKNOWLEDGMENTS

We thank Francisco Pereira for useful suggestions, and Steven Ibara, Wouter Kool, Janani Prabhakar, and Natalia Córdova for help with running participants. J.J.F.R.-F. was supported by the Fundação para a Ciência e Tecnologia, scholarship SFRH/BD/33273/2007, A.S. by an INRSA Training Grant in Quantitative Neuroscience 2 T32 MH065214, A.G.B. by AFOSR Grant FA9550-08-1-041, Y.N. by a Sloan Research Fellowship, and M.M.B. by the National Institute of Mental Health Grant P50 MH062196 and a Collaborative Activity Award from the James S. McDonnell Foundation.

Accepted: May 26, 2011
Published: July 27, 2011

REFERENCES

Badre, D. (2008). Cognitive control, hierarchy, and the rostro-caudal organization of the frontal lobes. Trends Cogn. Sci. (Regul. Ed.) *12*, 193–200.

Badre, D., and Frank, M.J. (2011). Mechanisms of hierarchical reinforcement learning in cortico-striatal circuits 2: evidence from fMRI. Cereb. Cortex, in press. Published online June 21, 2011.

Badre, D., Hoffman, J., Cooney, J.W., and D'Esposito, M. (2009). Hierarchical cognitive control deficits following damage to the human frontal lobe. Nat. Neurosci. *12*, 515–522.

Baker, T.E., and Holroyd, C.B. (2011). Dissociated roles of the anterior cingulate cortex in reward and conflict processing as revealed by the feedback error-related negativity and N200. Biol. Psychol. *87*, 25–34.

Barto, A.G. (1995). Adaptive critics and the basal ganglia. In Models of Information Processing in the Basal Ganglia, J.C. Houk, J. Davis, and D. Beiser, eds. (Cambridge, MA: MIT Press), pp. 215–232.

Barto, A.G., and Mahadevan, S. (2003). Recent advances in hierarchical reinforcement learning. Discrete Event Dyn. Syst. *13*, 341–379.

Botvinick, M.M. (2008). Hierarchical models of behavior and prefrontal function. Trends Cogn. Sci. (Regul. Ed.) *12*, 201–208.

Botvinick, M., Nystrom, L.E., Fissell, K., Carter, C.S., and Cohen, J.D. (1999). Conflict monitoring versus selection-for-action in anterior cingulate cortex. Nature *402*, 179–181.

Botvinick, M.M., Niv, Y., and Barto, A.C. (2009). Hierarchically organized behavior and its neural foundations: a reinforcement learning perspective. Cognition *113*, 262–280.

Brainard, D.H. (1997). The psychophysics toolbox. Spat. Vis. *10*, 433–436.

Breiter, H.C., Aharon, I., Kahneman, D., Dale, A., and Shizgal, P. (2001). Functional imaging of neural responses to expectancy and experience of monetary gains and losses. Neuron *30*, 619–639.

Christoff, K., Keramatian, K., Gordon, A.M., Smith, R., and Mädler, B. (2009). Prefrontal organization of cognitive control according to levels of abstraction. Brain Res. *1286*, 94–105.

Christoph, G.R., Leonzio, R.J., and Wilcox, K.S. (1986). Stimulation of the lateral habenula inhibits dopamine-containing neurons in the substantia nigra and ventral tegmental area of the rat. J. Neurosci. *6*, 613–619.

Cooper, R., and Shallice, T. (2000). Contention scheduling and the control of routine activities. Cogn. Neuropsychol. *17*, 297–338.

Cooper, J.C., and Knutson, B. (2008). Valence and salience contribute to nucleus accumbens activation. Neuroimage *39*, 538–547.

Corbetta, M., Patel, G., and Shulman, G.L. (2008). The reorienting system of the human brain: from environment to theory of mind. Neuron *58*, 306–324.

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

Daw, N.D., and Frank, M.J. (2009). Reinforcement learning and higher level cognition: introduction to special issue. Cognition *113*, 259–261.

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

Dietterich, T.G. (1998). The MAXQ method for hierarchical reinforcement learning. In Proceedings of the Fifteenth International Conference on Machine Learning, J.W. Shavlik, ed. (San Francisco: Morgan Kaufman), pp. 118–126.

Grafman, J. (2002). The human prefrontal cortex has evolved to represent components of structured event complexes. In Handbook of Neuropsychology, J. Grafman, ed. (Amsterdam: Elsevier).

Grent-'t-Jong, T., and Woldorff, M.G. (2007). Timing and sequence of brain activity in top-down control of visual-spatial attention. PLoS Biol. *5*, e12.

Hare, T.A., O'Doherty, J., Camerer, C.F., Schultz, W., and 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.

Haruno, M., and Kawato, M. (2006). Heterarchical reinforcement-learning model for integration of multiple cortico-striatal loops: fMRI examination in stimulus-action-reward association learning. Neural Netw. *19*, 1242–1254.

Hestvik, A., Maxfield, N., Schwartz, R.G., and Shafer, V. (2007). Brain responses to filled gaps. Brain Lang. *100*, 301–316.

Holroyd, C.B., and Coles, M.G.H. (2002). The neural basis of human error processing: reinforcement learning, dopamine, and the error-related negativity. Psychol. Rev. *109*, 679–709.

Holroyd, C.B., Nieuwenhuis, S., Yeung, N., and Cohen, J.D. (2003). Errors in reward prediction are reflected in the event-related brain potential. Neuroreport *14*, 2481–2484.

Holroyd, C.B., Nieuwenhuis, S., Yeung, N., Nystrom, L., Mars, R.B., Coles, M.G.H., and Cohen, J.D. (2004). Dorsal anterior cingulate cortex shows fMRI response to internal and external error signals. Nat. Neurosci. *7*, 497–498.

Jocham, G., and Ullsperger, M. (2009). Neuropharmacology of performance monitoring. Neurosci. Biobehav. Rev. *33*, 48–60.

Knutson, B., Taylor, J., Kaufman, M., Peterson, R., and Glover, G. (2005). Distributed neural representation of expected value. J. Neurosci. *25*, 4806–4812.

Kouneiher, F., Charron, S., and Koechlin, E. (2009). Motivation and cognitive control in the human prefrontal cortex. Nat. Neurosci. *12*, 939–945.

Krigolson, O.E., and Holroyd, C.B. (2006). Evidence for hierarchical error processing in the human brain. Neuroscience *137*, 13–17.

Lashley, K.S. (1951). The problem of serial order in behavior. In Cerebral Mechanisms in Behavior: The Hixon Symposium, L.A. Jeffress, ed. (New York: Wiley), pp. 112–136.

Matsumoto, M., and Hikosaka, O. (2007). Lateral habenula as a source of negative reward signals in dopamine neurons. Nature *447*, 1111–1115.

Miltner, W.H.R., Braun, C.H., and Coles, M.G.H. (1997). Event-related brain potentials following incorrect feedback in a time-estimation task: evidence for a "generic" neural system for error detection. J. Cogn. Neurosci. *9*, 788–798.

Montague, P.R., Dayan, P., and Sejnowski, T.J. (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. J. Neurosci. *16*, 1936–1947.

Nicola, S.M., Surmeier, J., and Malenka, R.C. (2000). Dopaminergic modulation of neuronal excitability in the striatum and nucleus accumbens. Annu. Rev. Neurosci. *23*, 185–215.

Nieuwenhuis, S., Slagter, H.A., von Geusau, N.J.A., Heslenfeld, D.J., and Holroyd, C.B. (2005). Knowing good from bad: differential activation of human cortical areas by positive and negative outcomes. Eur. J. Neurosci. *21*, 3161–3168.

Niv, Y. (2009). Reinforcement learning in the brain. J. Math. Psychol. *53*, 139–154.

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.

O'Doherty, J.P., Dayan, P., Friston, K.J., Critchley, H.D., and Dolan, R.J. (2003). Temporal difference models and reward-related learning in the human brain. Neuron *38*, 329–337.

O'Doherty, J.P., Buchanan, T.W., Seymour, B., and Dolan, R.J. (2006). Predictive neural coding of reward preference involves dissociable responses in human ventral midbrain and ventral striatum. Neuron *49*, 157–166.

Parr, R., and Russell, S. (1998). Reinforcement learning with hierarchies of machines. Adv. Neural Inf. Process Sys. *10*, 1043–1049.

Phan, K.L., Wager, T.D., Taylor, S.F., and Liberzon, I. (2004). Functional neuroimaging studies of human emotions. CNS Spectr. *9*, 258–266.

Poldrack, R.A. (2006). Can cognitive processes be inferred from neuroimaging data? Trends Cogn. Sci. (Regul. Ed.) *10*, 59–63.

Reynolds, J.R., and O'Reilly, R.C. (2009). Developing PFC representations using reinforcement learning. Cognition *113*, 281–292.

Salas, R., Baldwin, P., de Biasi, M., and Montague, P.R. (2010). BOLD responses to negative reward prediction errors in human habenula. Front. Hum. Neurosci. *4*, 36.

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

Seymour, B., O'Doherty, J.P., Dayan, P., Koltzenburg, M., Jones, A.K., Dolan, R.J., Friston, K.J., and Frackowiak, R.S. (2004). Temporal difference models describe higher-order learning in humans. Nature *429*, 664–667.

Seymour, B., Daw, N., Dayan, P., Singer, T., and Dolan, R. (2007). Differential encoding of losses and gains in the human striatum. J. Neurosci. *27*, 4826–4831.

Singh, S., Barto, A.G., and Chentanez, N. (2005). Intrinsically motivated reinforcement learning. In Advances in Neural Information Processing Systems 17: Proceedings of the 2004 Conference, L.K. Saul, Y. Weiss, and L. Bottou, eds. (Cambridge, MA: MIT Press), pp. 1281–1288.

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

Sutton, R.S., Precup, D., and Singh, S. (1999). Between MDPs and semi-MDPs: a framework for temporal abstraction in reinforcement learning. Artif. Intell. *112*, 181–211.

Talairach, J., and Tournoux, P. (1988). Co-Planar Stereotaxic Atlas of the Human Brain (New York: Thieme Medical Publishers, Inc.).

Ullsperger, M., and von Cramon, D.Y. (2003). Error monitoring using external feedback: specific roles of the habenular complex, the reward system, and the cingulate motor area revealed by functional magnetic resonance imaging. J. Neurosci. *23*, 4308–4314.

van Veen, V., Holroyd, C.B., Cohen, J.D., Stenger, V.A., and Carter, C.S. (2004). Errors without conflict: implications for performance monitoring theories of anterior cingulate cortex. Brain Cogn. *56*, 267–276.

Yacubian, J., Gläscher, J., Schroeder, K., Sommer, T., Braus, D.F., and Büchel, C. (2006). Dissociable systems for gain- and loss-related value predictions and errors of prediction in the human brain. J. Neurosci. *26*, 9530–9537.

Yamaguchi, S., Tsuchiya, H., and Kobayashi, S. (1995). Electrophysiologic correlates of visuo-spatial attention shift. Electroencephalogr. Clin. Neurophysiol. *94*, 450–461.

Yeung, N., Botvinick, M.M., and Cohen, J.D. (2004). The neural basis of error detection: conflict monitoring and the error-related negativity. Psychol. Rev. *111*, 931–959.

Yeung, N., Holroyd, C.B., and Cohen, J.D. (2005). ERP correlates of feedback and reward processing in the presence and absence of response choice. Cereb. Cortex *15*, 535–544.