Please cite this article in press as: Ribas-Fernandes et al., A Neural Signature of Hierarchical Reinforcement Learning, Neuron (2011), doi:10.1016/j.neuron.2011.05.042

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; 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

Please cite this article in press as: Ribas-Fernandes et al., A Neural Signature of Hierarchical Reinforcement Learning, Neuron (2011), doi:10.1016/j.neuron.2011.05.042

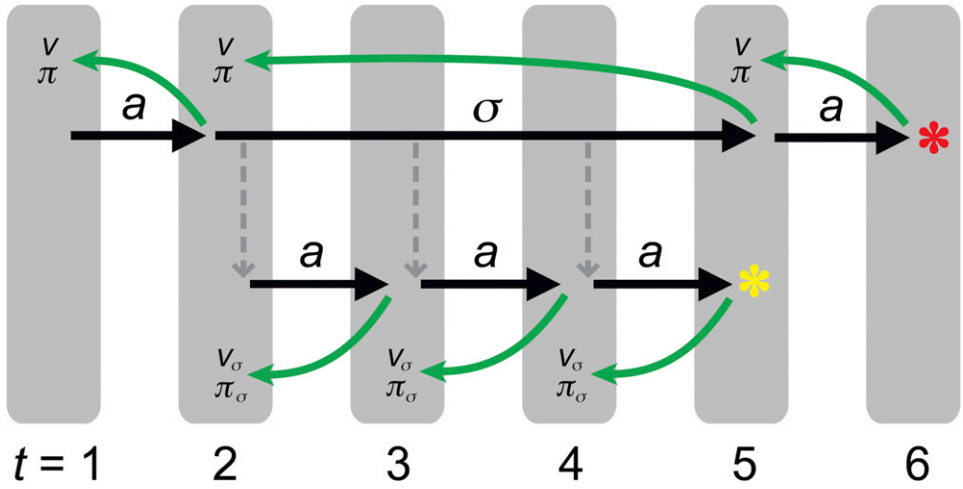


Figure 1. Illustration of HRL Dynamics

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

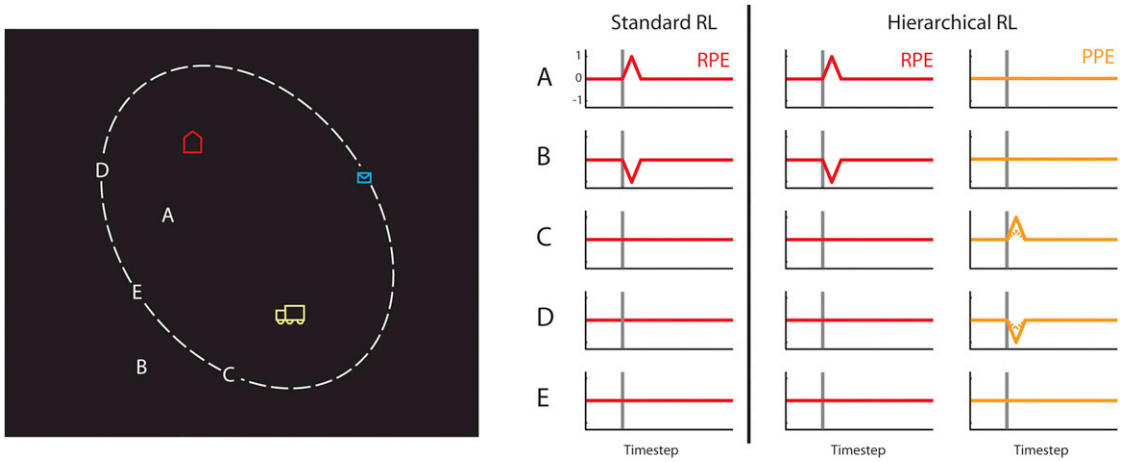


Figure 2. Task and Predictions from HRL and RL

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),

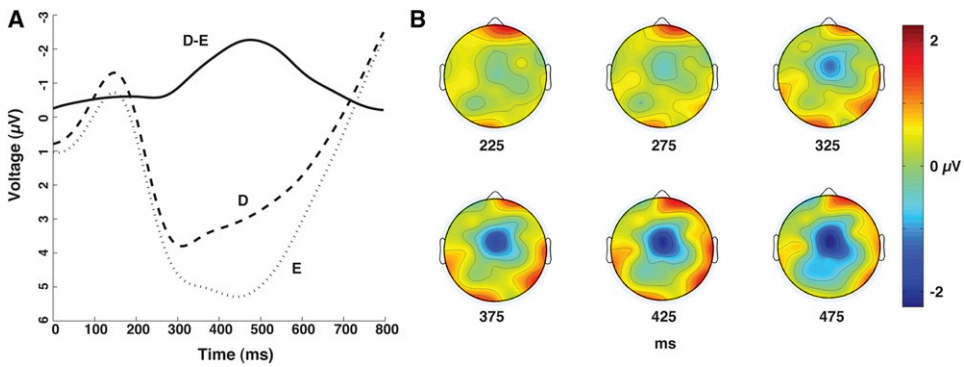


Figure 3. Results of EEG Experiment

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]).

Please cite this article in press as: Ribas-Fernandes et al., A Neural Signature of Hierarchical Reinforcement Learning, Neuron (2011), doi:10.1016/j.neuron.2011.05.042

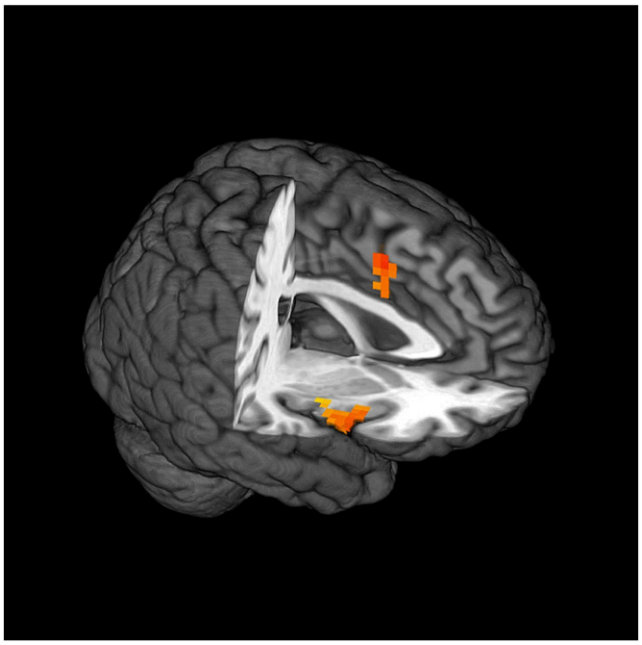


Figure 4. Results of fMRI Experiment 1

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.

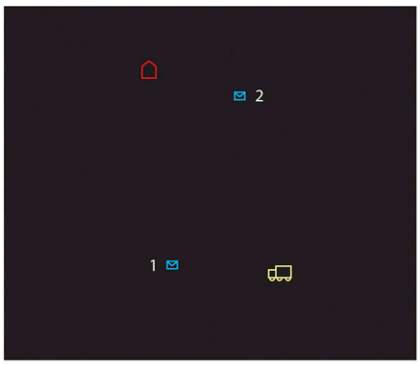
In a series of region-of-interest (ROI) analyses, we focused in on additional neural structures that, like the ACC, have been previously proposed to encode negative RPEs: the habenular complex (Salas et al., 2010; Ullsperger and von Cramon, 2003), nucleus accumbens (NAcc) (Seymour et al., 2007), and 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



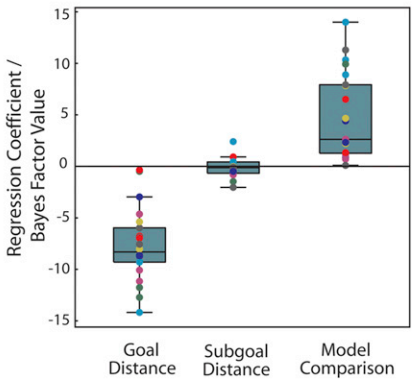


Figure 5. Results of Behavioral Experiment

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

Please cite this article in press as: Ribas-Fernandes et al., A Neural Signature of Hierarchical Reinforcement Learning, Neuron (2011), doi:10.1016/j.neuron.2011.05.042

[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; 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 (ρ) 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.

Diuk, C., Botvinick, M.M., Barto, A.G., and Niv, Y. (2010). Hierarchical reinforcement learning: an fMRI study of learning in a two-level gambling task. Society for Neuroscience Meeting, in press.

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.

Two simultaneous, but separable, prediction errors in human ventral striatum

Carlos Diuk, Karin Tsai, Jonathan Wallis, Matthew Botvinick and Yael Niv

In recent years, computational reinforcement learning (RL; Sutton & Barto, 1998) has provided an indispensable framework for understanding the neural substrates of learning and decision making. Dopaminergic signals projecting into the striatal nuclei, once elusive and misunderstood, are now widely thought to correlate with a scalar prediction error signal that indicates the difference between reward expectations and actual observations (Barto, 1995; D'Ardenne, McClure, Nystrom, & Cohen, 2008; Glimcher, 2011; Montague, Peter Dayan, & Sejnowski, 1996; W. Schultz, Peter Dayan, & Montague, 1997). This prediction error signal is key for learning about rewards in the world, and an indispensable element in RL models of learning.

While work studying midbrain function suggests that dopaminergic neurons all report one unitary prediction error signal (W. Schultz et al., 1997), recent physiological work suggests heterogeneity (Brischoux, Chakraborty, Brierley, & Ungless, 2009). Moreover, computational RL models that attempt to scale beyond simple action-outcome associations into real world tasks suggest that more than one prediction error may become necessary at the same time (Sutton, Precup, & Singh, 1999). Here we ask: Can the brain represent more than one reward prediction error at a time?

One case in which multiple, simultaneous reward prediction errors are needed is when performing tasks with hierarchical structure. This is because in hierarchical settings, the outcomes of multiple levels of a task structure might be observed at the same time, and the brain must update its expectations about each level separately. For example, imagine arriving at a city with multiple casinos, with a set of coupons that allow you to enter any of the casinos and play a number of different games. You enter one casino and play blackjack, roulette and a slot machine. Each time you play a game, you might observe a difference between what you expected to win, and the actual outcome – a "game-level prediction error" that can be used to adjust your future expectations from this game. However, upon playing the last coupon for this casino, you not only learn about this last game itself, but you also have enough information to update your knowledge about the casino as a whole: was this a good casino to squander your coupons in? It is at this point that you should use two coincident reward prediction errors: a game-related prediction error to learn the value of the last game, and a casino-related prediction error to learn the value of the casino as a whole, so that next time you use a set of coupons you can choose casinos wisely.[i]

Learning in this multi-level setting has been addressed in the computational framework of hierarchical reinforcement learning (HRL; Barto & Mahadevan, 2003; Dietterich, 2000; Sutton, Precup, & Singh, 1999). Recent work has suggested that

HRL may be relevant to human learning (Ribas-Fernandes et al., 2011), but did not address the key HRL prediction that simultaneous prediction error signals are necessary for learning at different levels of a hierarchy.

To test whether multiple distinct prediction errors exist in the human brain at the same time, we designed a task akin to the casino example above, essentially an extension of the classic bandits task to a hierarchical setting, and used fMRI to record BOLD signals while participants learned to play this task. We were specifically interested in BOLD signals in the ventral striatum, an area where activity has repeatedly been shown to correlate with prediction error signals (Glimcher, 2011; Hare, John P. O'Doherty, Camerer, W. Schultz, & Rangel, 2008; Y. Niv, Edlund, P. Dayan, & J. P. O'Doherty, 2012). To model learning in this setting, we used the computational framework of HRL.

Results

Twenty-eight participants played 120 trials of a two-level "casino" task, modeled after the scenario described in the Introduction, while in an MRI scanner (see Experimental Procedures). In each trial, participants first chose between two doors, representing two casinos. Once a casino was chosen, its door opened and a "target" was revealed – a number of points (2 to 10, distributed normally with means 5 and 6 in each of the two casinos) that must be accumulated in order to gain a reward of 10 cents in the casino. Each casino also contained a unique set of 4 slot machines, of which participants chose two to play. Each slot machine granted 0-5 points, normally distributed, with an independent, slowly drifting mean. If they did not succeed in meeting the target with their two plays, participants lost 10 cents.

This task was designed to elicit learning at two levels: at the slot machine level (to inform choices within a casino) and at the casino level (to inform choices between the two casinos). In particular, two distinct and coincident prediction errors should occur after playing the second slot machine, when the point outcome of that machine is revealed simultaneously with the win/lose 10¢ outcome of the casino as a whole. Importantly, in this design these two prediction errors are uncorrelated: it is possible to obtain fewer points than expected on the second slot machine (a negative slot-level prediction error) while at the same time still win the casino as a whole (a positive casino-level prediction error), and vice versa.

Behavioral results: Subjects learn at multiple levels of the hierarchy

We first sought to verify that participants were learning at both levels of the task, that is, we tested whether they learned to choose both the best casino as well as the best slot machines within the chosen casino. Note that the best casino is the one with the highest expected probability of winning, a probability that depends both on the distribution of the casino's point targets and the expected points that its slot machines will yield. For instance, one casino might tend to have lower targets,

making it initially more attractive to a naive player, but its slot machines might tend to yield few points, resulting in an overall lower probability of winning. Moreover, the true mean number of points for each slot machine at each trial is unknown and can only be estimated through experience with the slot machine. To evaluate how good the participants' sequences of choices were, we had to take into account the information that they could have gleaned from previous samples of each slot machine and use a model-based analysis. To test whether participants correctly integrated information from both levels in order to determine the value of each casino, we fit two learning models to participants' behavioral choices at the casino level: 1) A correct *Outcome Model*, that assumes that the participant updated the expected value of a casino based on the casino's true outcomes after playing the two slot machines, using a standard temporal-difference learning mechanism; and 2) A more straightforward, but incorrect, alternative *Target Model*, that assumes that the participant updated the expected values of the casinos based only on the point target of the casino (see *Experimental Procedures* for details of both models). Formal model comparison of the *Outcome* vs. *Target* models showed overwhelming support for the *Outcome* model, which better fit the choices of 22 of the 28 participants (Figure 1; $p<0.002$, one-tailed paired Student's t-test on the difference in log-likelihoods of the *Outcome* model and the *Target* model).

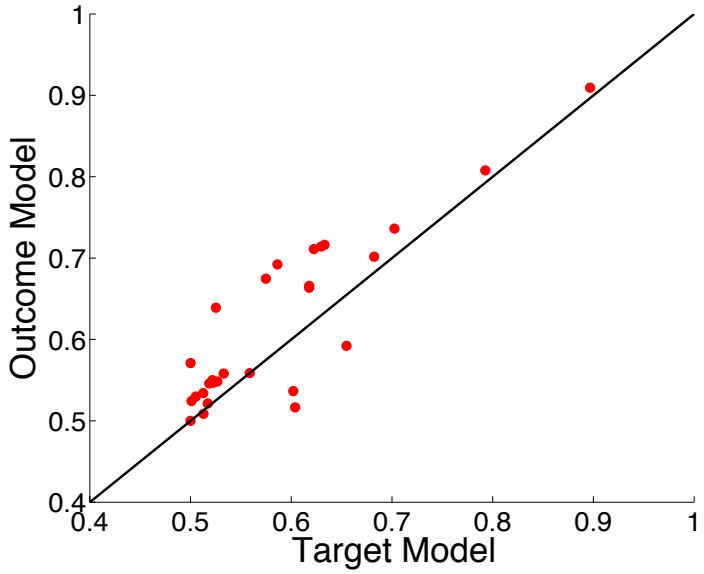


Figure 1. Average posterior probability per choice trial for the *Outcome* and *Target* models, per participant. The *Outcome* model assigns a higher average probability per trial to the choices of 22 out of 28 subjects (points lying above the solid equal-likelihood line). The average probability of a choice trial was calculated as the likelihood of the whole sequence of choice data divided by the number of choice trials.

To verify that participants learned at the slot machine level, we first compared their actual slot machine payoffs to those of a simulated agent that chose slot machines randomly. The agent was presented with the series of casinos that the participant chose, using the same slot machine point distributions, and chose two of the machines at random. A comparison of the total number of slot machine points accumulated by the random agent, averaged over 100 random agents per participant, and the actual points obtained by the participant showed that participants earned significantly more points than random agents ($p<10^{-5}$, one-tailed paired Student's t-test).

This result shows that participants are learning about slot machines. It still does not show that they are learning based on individual slot-level prediction errors. To test this hypothesis, we compared two alternative learning models: 1) A *slot-points model* that assumes that participants chose which two slot machines to play based on the sum of their expected outcomes, and updated the expected value of each slot machine after observing its outcome; and 2) a *6-armed bandit model* that treats each of the 6 possible pairs of slot machines as a different "arm" in a bandit problem, with the value of the "arm" updated at the end of the trial based on the overall casino outcome. The first model represents a standard TD learning model, but accounting for the fact that the order of the two chosen slot machines is inconsequential. The second model also uses TD learning, but does not learn specifically about the slot machine outcomes. That is, rather than learn from slot-machine-specific prediction errors at the lower level of the hierarchy, it learns a choice policy based on the casino outcomes. Here, again, formal model comparison favored the *slot-points* model that learns from slot-machine-specific prediction errors: the *slot-points* model better fit the choices of 25 of the 28 participants, with the remaining three participants equally fit by the two models (Figure 2; $p<10^{-6}$, one-tailed paired Student's t-test on the difference in log-likelihoods of the *slot-points* model and the *6-armed-bandit* model).

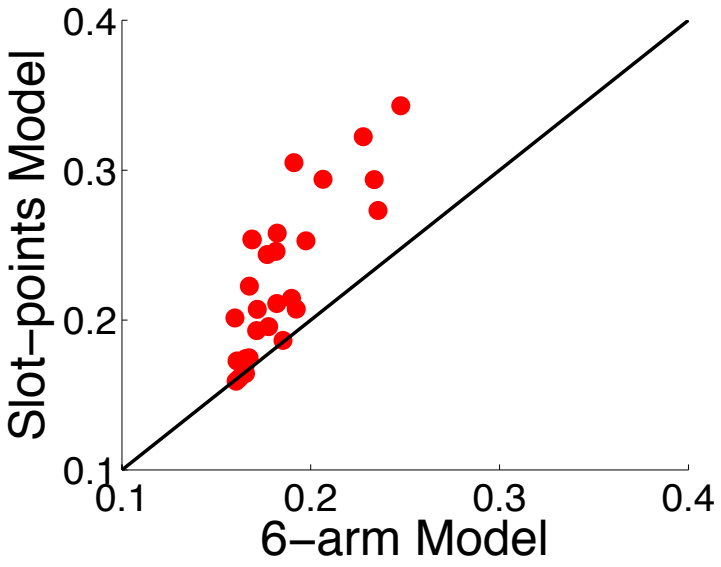


Figure 2. Average posterior probability per choice trial for the *TD-pairs* and *6-armed bandit* models, per participant. The *TD-pairs* model assigns a higher average probability per trial to the choices of 25 out of 28 subjects (points lying above the solid equal-likelihood line).

Together, these results suggest that subjects were simultaneously learning about slot machines and casinos in a temporal-difference fashion based on separate prediction errors at each level of the hierarchy. Thus we should expect concurrent, distinct prediction errors, at least at the time of the last slot machine play, at which point information about that slot machine *and* about the overall worth of the casino became available simultaneously.

fMRI results: Two concurrent prediction errors in ventral striatum

Based on the extensive existing literature on BOLD correlates of prediction error signals in the human brain, we focused our analysis on an *a priori* anatomically-defined region of interest (ROI) in the ventral striatum (VS) (Abler, Walter, Erk, Kammerer, & Spitzer, 2006; Delgado, Miller, Inati, & Phelps, 2005; Glimcher, 2011; Hare et al., 2008; Li, McClure, King-Casas, & Read Montague, 2006; McClure, Berns, & Montague, 2003; Y. Niv et al., 2012; John P. O'Doherty, Peter Dayan, Friston, Critchley, & Dolan, 2003; John P. O'Doherty et al., 2004; Preuschoff, Bossaerts, & Quartz, 2006), delineated separately for each participant using their structural brain image. Data were extracted from this ROI and averaged using singular value decomposition, resulting in one time-course vector of VS BOLD activity per participant (see *Experimental Procedures*).

Supported by the behavioral results described above, we used the framework of hierarchical reinforcement learning (HRL) to model the participants' learning and choice behavior (see *Experimental Procedures* for details) and to generate

regressors for analysis of the fMRI data. At the lower slot-machine level, our model included a standard TD learning mechanism that learns a separate value for each slot machine, updating the value estimate every time the outcome of this machine is encountered according to the difference between the expected value and the actual number of points obtained as in the *slot-points* model above. At the higher casino level, a separate TD learning mechanism kept track of the value of each casino, updating it at the time of the casino outcome (as in the previously described *Outcome* model). This hierarchical model thus induced three regressors of interest in each trial, for each participant: 1) a prediction error for the first slot machine played ("FirstSlot"); 2) a prediction error for the second slot machine ("LastSlot"); and, 3) a prediction error for the casino ("Casino"). We modeled each regressor at the onset of the outcome that led to that prediction error, with LastSlot and Casino occurring simultaneously. Through the design of the task, these regressors were nearly orthogonal (mean correlation coefficient -0.029, std 0.07) allowing us to search for neural correlates of each despite their temporal co-occurrence.

All three regressors of interest were significantly correlated with VS BOLD activity (FirstSlot, $p<0.004$; LastSlot, $p<0.0269$; Casino, $p<5.5x10^{-5}$), indicating that indeed two distinct, but temporally coincident prediction error signals, LastSlot and Casino, can co-exist in the VS. To verify the result from our behavioral model comparison, we also tested whether the VS signal correlated with a prediction error regressor based on target only (from the *Target* model), and found no significant effect ($p=0.22$).

Whole brain analysis

Our previous analysis concentrated on *a priori*, anatomically defined ROIs in the ventral striatum. To supplement this, we conducted a whole brain analysis, searching for areas correlating with two regressors of interest (FirstSlot+LastSlot combined and Casino). At a whole-brain corrected threshold of $p>0.05$, the only significant positive correlation found was between the Casino regressor and bilateral ventral striatum (Figure 2). We also observed large clusters negatively correlated with the Casino regressor in visual areas (see Table 1). It is worth noting that when participants win the Casino, the points bar turns green, whereas when they lose part of the bar is yellow (indicating the points they did achieve) and the number of points by which they missed the target is indicated in red. While the origins of these negative correlations are unclear, we speculate that when participants lose in the Casino (resulting in a negative PE and a higher visual activation), there is an increase in attentive visual processing. Potentially, participants more attentively look at the number of points they got and by how much they missed, while not paying so much attention at wins.

Anatomical location	**Peak x,y,z (mm)**	**Cluster Size**	**Peak intensity (T)**
Right ventral putamen	18,14,-8	29	7.15

Left ventral putamen	-18,8,-11	17	7.05
Left Lingual Gyrus	-3,-64,1	31	-6.91
Right Lingual Gyrus	9,-73,-5	7	-6.66
Occipital lobe	-12,-88,25	485	-12.68

Table 1. All activations that survived a whole-brain FWE corrected threshold of $p<0.05$ in the random effects contrast for a Casino prediction error signal. Anatomical locations were determined through inspection with respect to the average anatomical image of all 28 participants.

The combined slot machine regressor did not reveal any activations that survived whole brain correction. However, bilateral ventral striatum activations did survive the commonly reported uncorrected $p<0.001$ threshold, as expected (Figure 2). The relatively weaker activations for the slot machine regressor, as compared to the Casino regressor, were likely due to the fact that outcomes at the casino level were binary, while slot machine outcomes were normally distributed, and as a result the variance of the Casino regressor was an order of magnitude larger than that of the slot machine regressors. This speculation is in line with previous work that suggests that prediction errors due to binary outcomes are easier to detect in ventral striatal BOLD signals, than those derived from normally-distributed outcomes (compare, for instance, Figure 2a in (Schönberg, Daw, Joel, & John P. O'Doherty, 2007), to the activations in supplemental table 3 in (Daw, John P. O'Doherty, Peter Dayan, Seymour, & Dolan, 2006). We verified this hypothesis in a companion experiment (not reported here) in which slot machine outcomes were binary (win/lose) and the casino outcome was multiple-valued (specifically, participants could play up to four slot machines each costing 5¢ to play, and had to win at least two slot machines to earn a certain amount of money the casino offered, which varied in each trial according to a casino-specific distribution). In this companion experiment, whole-brain analysis (FWE corrected at the $p<0.05$ level) revealed significant bilateral VS activation for the slot-machine prediction-error regressors, and no significant activation for the casino regressor. We could not design a task in which both levels resulted in binary outcomes due to the constraint of orthogonalizing LastSlot outcomes and Casino outcomes (which was also not met by the design of the companion experiment).

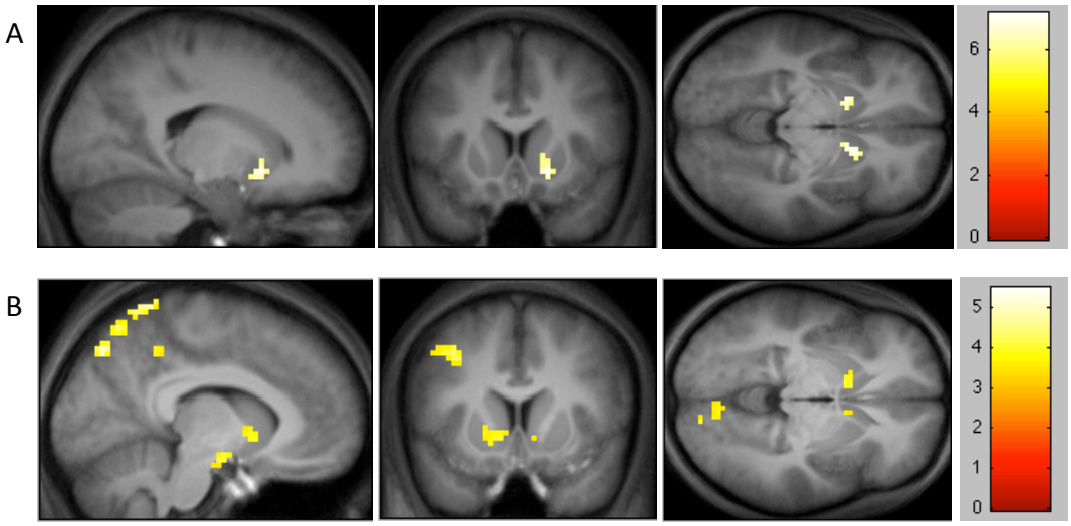


Figure 3. A) Activations that survived a whole-brain FWE corrected threshold of p<0.05, cluster size>5, in the random effects contrast for the Casino regressor. Images are centered at voxel (18,14,-8). B) Activations that survived an uncorrected threshold of p<0.001, cluster size>5, in the random effects contrast for the combined Slot regressor. Images are centered at voxel (-9,11,-5).

Anatomical separation between Slot and Casino activations

The presence of two simultaneous prediction error signals leads to a natural question: is there some anatomical separation within striatum between areas activated by the slot machine prediction errors and areas activated by casino prediction errors? We divided the striatum into three anatomical components, at an individual subject level: ventral striatum (vStr), Putamen and Caudate. Activity in vStr and Caudate correlates with all three regressors (FirstSlot, LastSlot and Casino), whereas activity in Putamen only correlates with Casino and FirstSlot, but not LastSlot. Based on the fact that activations for LastSlot, which coincide in time with Casino, are weaker than those of FirstSlot, we find this result still inconclusive and the object of further research. As an extra element to study anatomical separation, we plotted the mean regression coefficients, across subjects, within a group-level vStr ROI as slices along the y coordinate (Figure 4). Although the pattern of activations does not lead to any statistically conclusive result, simple observation shows a structure of stripes, where slot regressors being more active medially and the casino regressor more lateral.

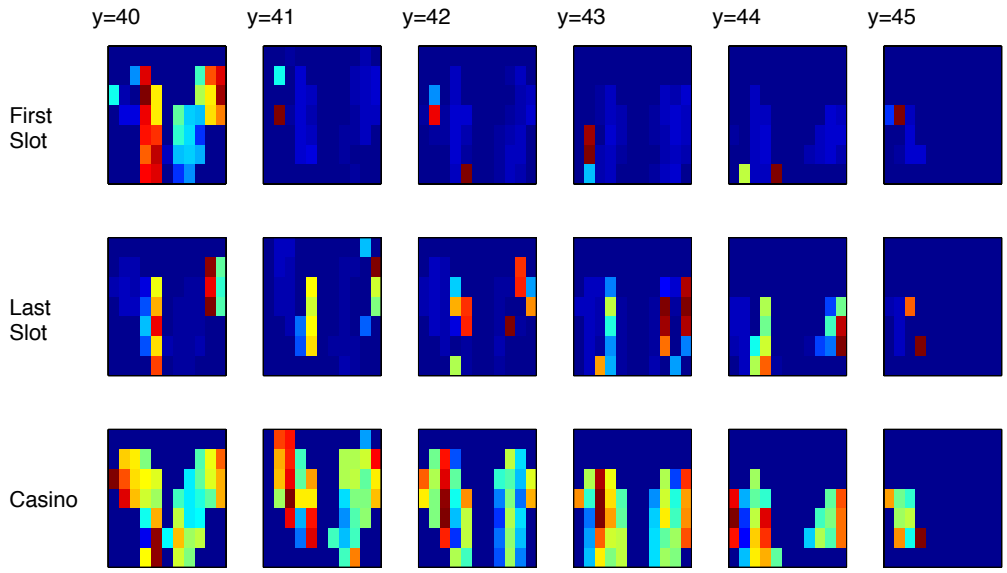


Figure 4. Mean regression coefficients per voxel, for each regressor, within a group-level vStr ROI.

Discussion

We investigated whether, when confronted with a hierarchical task that requires learning about more than one level of the hierarchy at the same time, the brain is capable of generating two simultaneous but distinct reward prediction error signals. For this we designed a task that, under our HRL model, involves two simultaneous but uncorrelated prediction error signals. Behavioral results showed that participants learned the task successfully at both levels of the hierarchy. Moreover, they showed that participants used the outcome of the low-level sub-tasks (slot machines) and of the high-level task (casino) to make future choices. The latter outcome was only revealed at the end of the high-level task, at the exact same time as the outcome of a lower-level task (slot-machine play). Learning about these two separable but coincident events thus required the presence of two simultaneous, distinct reward prediction errors. Using fMRI and anatomically defined ROIs in the ventral striatum, we found that BOLD signals in the ventral striatum indeed correlated with both these prediction error signals.

Our work adds to a growing body of evidence indicating the existence of more than one prediction error signal in the brain. Neuroimaging data has suggested the existence of functionally distinct error signals, namely a reward and a state prediction, albeit in different brain areas (Gläscher, Daw, Dayan, & O'Doherty, 2010). Physiological findings have also suggested that the population of dopaminergic neurons may not be homogenous (Brischoux et al., 2009). Our results extend beyond these previous reports by showing that reward prediction error signals projecting into the same brain area (ventral striatum) can signal more than one

quantity at the same time.

Our results thus provide functional evidence against the unitary nature of prediction errors, and suggest that in more realistic learning scenarios several prediction errors may be used to learn in parallel at different levels of the task. This finding contrasts with previous empirical work suggesting that prediction errors are scalar and unitary, resulting from dopamine neurons signaling a single, global difference between obtained and expected reward (Glimcher, 2011; W. Schultz et al., 1997). However, it does not contradict those previous data: in the simple tasks previously examined, only one prediction error signal was available and required for learning at each point in time. One recent study (Daw, Gershman, Seymour, Peter Dayan, & Dolan, 2011) indicates the presence of a striatal prediction error signal based on combined predictions from two learning systems (model-based and model-free). In this work, however, the two signals were not orthogonal and only their combined effect could be observed. Our finding of two concurrent prediction error signals suggests that these prior results are a special case of the function of prediction errors in the ventral striatum, and that a more detailed parcellation can be uncovered by using more complex tasks. Indeed, our results suggest that the so-called 'scalar prediction error signal' may be more of a vector-valued signal, as required by a number of reinforcement learning extensions like learning of successor representations (Peter Dayan, 1993; Hayes, Petrov, & Sederberg, 2011), factored representations (Koller & Parr, 1999) and hierarchical reinforcement learning (Barto & Mahadevan, 2003; Botvinick, Yael Niv, & Barto, 2009). The result of our work raises the problem of credit assignment: how does the brain distinguish the origin of the different signals and decomposes the vectorized prediction error in order to learn about different entities separately. We believe this question to be key in future research.

To generate two simultaneous prediction errors, we made use of hierarchy and predictions from the computational framework of HRL (Barto & Mahadevan, 2003; Dietterich, 2000; Sutton et al., 1999). Despite differences between existing HRL models, common to most of them is the existence of multiple prediction errors occurring when a sub-task ends and new knowledge about multiple levels of the hierarchy becomes available simultaneously. Recent work (Botvinick et al., 2009) derived a set of predictions from the HRL framework, evaluating the extent to which current scientific knowledge accorded with each of its elements. Only recently, experimental data has been produced to directly support some of these predictions (Badre & Frank, 2011; Ribas-Fernandes et al., 2011). Our work provides evidence for a key prediction of HRL, namely the existence of coincident reward prediction errors corresponding to the different levels of the hierarchy, and thus further supports the potential relevance of HRL to human behavior and brain function.

Experimental Procedures

Participants. 30 participants were recruited from the University community, and

gave informed consent. Two participants were excluded due to technical problems, and all data analysis was performed on the remaining 28 (ages 18-38, mean 22.04, 13 males, all right-handed). Participants received monetary compensation of $20 per hour, plus a small monetary bonus based on task performance (participants began the task with a budget of $1 and kept any money earned in the casinos, resulting in average earnings of $2.34, std=1.39, min=-0.45, max=4.55). All experimental procedures were approved by the Institutional Review Board of Princeton University.

Task and Procedure. The computerized task was coded using Matlab (The Mathworks, Natick, MA) using the Psychophysics toolbox, version 3 (Brainard, 1997). Participants played 120 trials split into 4 blocks of 30 trials each. Between blocks, they were given the option to take a break. On each trial, two doors representing the two different casinos appeared (Figure 3). Each door had a sign on it that said "Open" or "Closed". In 70 of the 120 trials, both doors were "Open", and in 50 trials (randomly interspersed among the 120) one of them was "Closed", forcing participants to choose the only open casino. Participants used a response trigger box to select one of the open casinos. Once participants chose a casino, its door opened revealing a bar that graphically indicated, in red, how many points needed to be accumulated in order to win 10¢. The number of points needed to win (target points) was drawn from a Normal distribution with mean 5 for the left casino and 6 for the right casino, and a standard deviation of 2.5 in both cases. The resulting draw was rounded to the closest integer and a minimum of 2 and a maximum of 10 were imposed.

After a jittered time interval that lasted between 2.5 and 3.5 seconds, four slot machines were displayed inside the corresponding casino, each a different and unique color. Participants played the game by selecting a slot machine with one of the four buttons in the trigger box. When a machine was selected, the other 3 machines were temporarily deactivated (graphically depicted by turning gray). The selected machine was animated to simulate spinning for 200ms and then displayed a number of points obtained. The number of points was shown as a green bar inside the slot machine, with a roman numeral to its side. The top bar indicating the target points was also updated in the following way: if enough points were accrued to win, the bar turned green. If not, the portion of the total target points just accrued in the slot machine became yellow, with the remaining bar still red. Each machine could produce a number of points drawn from a Normal distribution; with the drawn number rounded to the closest integer and bounded between 0 and 5. Each of the 8 slot machines (4 per casino) followed a different Normal distribution, with its own randomly drifting mean and a standard deviation of 1. The mean number of points of each slot machine drifted after each trial by +0.5 or -0.5 (drawn randomly with equal probability).

After the first slot machine play, a jittered wait time of 2.5 to 3.5 was imposed until the inactive machines became active again and a second slot machine could be chosen. After the second slot machine was selected, it spun for another 200ms and

displayed its number of accrued points. Once again, the target-points bar added the points just accrued and turned green in case of a win, or stayed partially yellow/partially red if not. A message was also displayed indicating that the trial was over and the amount of money earned: "Exiting Casino. Total Earned: +/-10¢". After a jittered wait time of 2.5 to 3.5 seconds, the door closed and a new trial began.

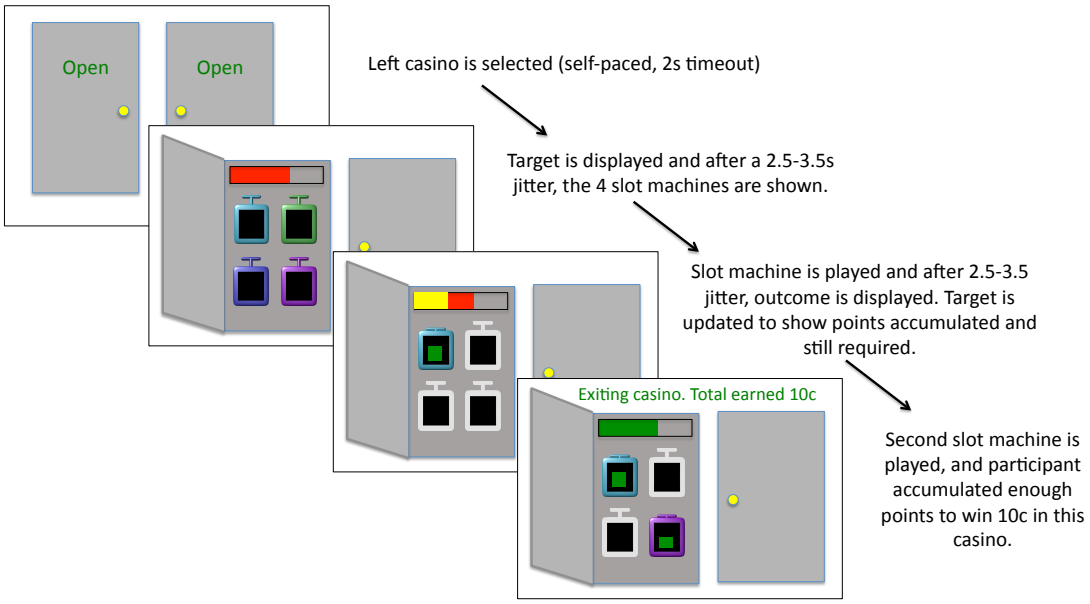


Figure 3. Sample trial: participant chooses to play in the left casino, the door opens and displays a target number of points (indicated by red bar). After a few seconds, the four slot machines appear. Participant plays upper-left slot and after a few seconds, the points obtained in that machine are shown inside the machine (green bar). Part of the target turns yellow, indicating the points accumulated with the first slot machine play. The rest is still red, indicating the points still necessary to win the casino. Participant plays the bottom right slot machine and obtains enough points to win the casino (10¢). The target bar turns green and a message appears indicating the casino win.

Behavioral data fits and model comparison

Temporal difference (TD) reinforcement learning (Sutton & Barto, 1998) provides a general framework for understanding trial-by-trial learning and decision making in simple tasks. A number of extensions have been proposed in the computational literature to the case where tasks involve hierarchical structure (Barto & Mahadevan, 2003; Sutton et al., 1999). What all of these extensions have in common is that they introduce the notion of temporally extended action sequences. The implication for human behavior is that some decisions no longer involve a single time-step, but commit the behaving agent to a longer set of steps. Variants of TD learning have been proposed to the temporally extended setting, and we use them to model the Casino task presented in this paper.

Playing a casino is a temporally extended action, which involves a series of slot machine plays. Following the "options" model from HRL (Sutton et al., 1999), we assume that a value V_{casino} is maintained for each of the casinos. We posit two different models for what these values represent and how they are updated:

a) The "*Target model*" posits that casino values are solely based on the target number of points required to win in them. This simple model ignores the fact that the quality of a casino also depends on the expected quality of its slot machines, and, in a sense, ignores the hierarchical nature of the task. Under the target model, as soon as the casino door opens, the target points p(t) are revealed and a prediction error $\delta(t) = p(t) - V(t)$ is computed. The value of the casino is updated at this time based on $V^{new}_{casno(i)} = V^{old}_{casno(i)} + \eta\delta(t)$.

b) The *"Outcome model"* posits that casino values are based on the probability of winning 10¢ in them. This value is therefore based on both the target points and the quality of the slot machines in the casino (as well as the policy used to play the slot machines). This model is, in that sense, fully hierarchical. As soon as the second slot machine is played, the participant observes a casino-level outcome $r_{casino}(t)$, which is either +10¢ or -10¢. A prediction error $\delta(t) = r_{casno}(t) - V(t)$ is computed and the value of the casino is updated based on $V^{new}_{casno(i)} = V^{old}_{casno(i)} + \eta\delta(t)$.

For the case of learning about slot machines, we assumed two different TD learning mechanisms:

a) The "*slot-points model*" posits that separate values $V_{slot(i)}$ are maintained for each of the eight slot machines in the game. At the beginning of the trial, participants choose a pair of slot machines *i* and *j* based on the sum of their values, so that $V_{pair(i,j)} = V_{slot(i)} + V_{slot(j)}$. This effectively ignores the order in which each slot is played within a Casino. The values for the individual slot machines are updated after their outcome is observed according to the simple TD rule $V^{new}_{slot(i)} = V^{old}_{slot(i)} + \eta\delta(t)$, with η being a learning rate or step-size parameter.

b) The *"6-armed bandit model"* posits that a value $V_{pair(i,j)}$ is maintained for each possible pair of slot machines. Since there are 4 slot machines per casino, there are 6 possible pairs. Participants use these pair values directly to choose which 2 slot machines they want to play. Rather than attending to the number of points obtained in each machine, they register whether or not they won or lost in the casino. The value of the pair of slot machines played is updated based on the TD rule $V^{new}_{pair(i,j)} = V^{old}_{pair(i,j)} + \eta\delta(t)$, with η being a learning rate or step-size parameter and $\delta(t)$=1 if they won, and $\delta(t)$=-1 otherwise.

For the slot machine choices as well as for the casino choices in both models, we assumed a soft-max action selection function:

$$p(A) = \frac{e^{\beta V(A)}}{\sum_{j \in actions} e^{\beta V(j)}}$$

where p(A) is the probability of choosing action A (a slot machine, or a casino), β is an inverse temperature parameter, and *j* enumerates all currently possible actions at this level of the hierarchy.

We used each participant's behavioral data to fit the models' free parameters η and β for the slot machines and, separately, another η and β for each of the two casino-learning models. Model likelihoods were based on assigning probabilities to each choice for each subject, according to the soft-max function specified above. In the case of the slot machines, there were 120 choices of pairs. In the case of the casinos, likelihood was only estimated based on the 70 choice trials, although we modeled learning of casino values using all 120 trials (note that each level of our task could be fit independently as, given the actual slot machine choices, there was no interaction between slot machine values and learning at the casino level).

We optimized model parameters by minimizing the negative log likelihood of the data given different parameter settings, using Matlab's fmincon function. This function performs constrained linear optimization over the space of possible parameter values. We provided the constraints on these values, setting η between 0 and 1, and β between 1 and 30. To facilitate finding the global minimum of the negative log likelihood, we ran the routine 4 times from different, randomly chosen starting initial values for these parameters, and kept track of the best fit over all runs. We compared the two alternative casino models by comparing the likelihoods of the models directly. Note that both models have the same number of parameters, so no penalties needed to be established.

fMRI data acquisition and preprocessing

Functional brain images were acquired using 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 (3x3x3 mm voxels, 41 contiguous 3 mm thick slices aligned with the anterior commissure - posterior commissure plane, interleaved acquisition, TR 2400 ms, TE 30 ms, flip angle 90°, field of view 192 mm).

Preprocessing of the images and whole brain image analysis were performed using SPM8 (Wellcome Department of Imaging Neuroscience, Institute of Neurology,

London, UK). Preprocessing of EPI images included motion correction (rigid body realignment of all images to the first volume), and spatial normalization to a standard T2* template in Montreal Neurological Institute (MNI) space. Anatomical ROIs were marked for each subject using MRIcron (Center for Advanced Brain Imaging, Georgia State and Georgia Tech Universities, Georgia, USA). Whole brain images were then further preprocessed by spatially smoothing the images using a Gaussian kernel with a full width at half maximum of 8mm, to allow for statistical parametric mapping analysis.

Region of Interest (ROI) analysis

The nucleus accumbens (NAC) was anatomically defined as the area bordered ventrally by the caudate nucleus, dorsally by the anterior commisure, laterally by the globus pallidus and putamen, and medially by the septum pellucidum. The border with the caudate was taken to be at the bottom of the lateral ventricle, and with the putamen at the thinnest part of grey matter. We considered the anterior-most border to be at the axial slice in which the caudate and putamen fully separated, and the posterior border where the anterior commisure was fully attached between hemispheres. Only voxels wholly within these boundaries were considered part of the ROI.

To analyze ROI timecourses, we first averaged, for each subject, the BOLD signal in all NAC voxels using single value decomposition. This resulted in a single timecourse per subject. We then removed from the timecourses effects of no interest due to scanner drift and participant motion by estimating and subtracting from the data, for each session separately, a linear regression model that included the six motion regressors (3D translation and rotation), two trend regressors (linear and quadratic) and a baseline. To test whether the resulting signal corresponded to prediction error signals, we regressed against each ROI timecourse a linear model which included three regressors of interest: FirstSlot, LastSlot and Casino, as described in the Results section. To conclude that BOLD activity corresponds to a prediction-error regressor, we required significant correlations at $p<0.05$ across subjects, and that these correlations be positive.

Whole-brain analysis

We used SPM8 to conduct a supplementary whole-brain analysis in which we searched for brain areas in which BOLD activity correlates with prediction error signals induced by our models. The design matrix comprised, for each of the four sessions: two parametric regressors for Slot PEs and Casino PEs according to the outcome model; two stick-function regressors for stimulus and outcome onsets (casino door opens, and each of the slot machine outcomes); and nuisance covariate regressors for motion, linear and quadratic drift, and baseline. The prediction error regressors were added as covariate regressors by convolving the punctuate prediction errors as assigned by the model with the canonical hemodynamic response function (HRF). Other stick regressors were convolved with the HRF as in

usual in SPM8. The six scan-to-scan motion parameters produced during preprocessing were used as nuisance motion regressors, to account for residual effects of movement. This design matrix was entered into a regression analysis of the fMRI data of each subject. A linear contrast of regressor coefficients was then computed at the single subject level for each regressor of interest. The results were analyzed as random effects at a second, between-subjects level, by including the contrast images of each subject in a one-way ANOVA with no mean term. Group level activations were localized using group-averaged structural scans using xjView (http://www.alivelearn.net/xjView).

References

Abler, B., Walter, H., Erk, S., Kammerer, H., & Spitzer, M. (2006). Prediction error as a linear function of reward probability is coded in human nucleus accumbens. *NeuroImage*, *31*(2), 790-795.

Badre, D., & Frank, M. J. (2011). Mechanisms of Hierarchical Reinforcement Learning in Cortico-Striatal Circuits 2: Evidence from fMRI. *Cerebral cortex (New York, N.Y. : 1991)*, 1-10. doi:10.1093/cercor/bhr117

Barto, A. G. (1995). Adaptive Critics and the Basal Ganglia. In J. C. Houk, J. Davis, & D. Beiser (Eds.), *Models of Information Processing in the Basal Ganglia* (pp. 215-232). Cambridge, MA: MIT Press.

Barto, A. G., & Mahadevan, S. (2003). Recent Advances in Hierarchical Reinforcement Learning. *Discrete Event Dynamic Systems*, *13*(4), 341-379. Hingham, MA, USA: Kluwer Academic Publishers. doi:http://dx.doi.org/10.1023/A:1025696116075

Botvinick, M. M., Niv, Yael, & Barto, A. G. (2009). Hierarchically organized behavior and its neural foundations: a reinforcement learning perspective. *Cognition*, *113*(3), 262-80. Elsevier B.V. doi:10.1016/j.cognition.2008.08.011

Brischoux, F., Chakraborty, S., Brierley, D. I., & Ungless, M. a. (2009). Phasic excitation of dopamine neurons in ventral VTA by noxious stimuli. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(12), 4894-9. doi:10.1073/pnas.0811507106

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, Peter, & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, *69*(6), 1204-15. Elsevier Inc. doi:10.1016/j.neuron.2011.02.027

Daw, N. D., O'Doherty, John P., Dayan, Peter, Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, *441*(7095), 876-879.

Dayan, Peter. (1993). Improving Generalization for Temporal Difference Learning: The Successor Representation. *Neural Computation*, *5*(4), 613-624. doi:10.1162/neco.1993.5.4.613

Delgado, M. R., Miller, M. M., Inati, S., & Phelps, E. a. (2005). An fMRI study of reward-related probability learning. *NeuroImage*, *24*(3), 862-73. doi:10.1016/j.neuroimage.2004.10.002

Dietterich, T. G. (2000). Hierarchical Reinforcement Learning with the MAXQ Value Function Decomposition. *Journal of Artificial Intelligence Research*, *13*, 227-303.

D'Ardenne, K., McClure, S. M., Nystrom, L. E., & Cohen, J. D. (2008). BOLD Responses Reflecting Dopaminergic Signals in the Human Ventral Tegmental Area . *Science* , *319* (5867), 1264-1267.

Glimcher, P. W. (2011). Understanding dopamine and reinforcement learning: The dopamine reward prediction error hypothesis. *Proceedings of the National Academy of Sciences of the United States of America*, *2010*. doi:10.1073/pnas.1014269108

Hare, T. a, O'Doherty, John P., Camerer, C. F., Schultz, W., & Rangel, A. (2008). Dissociating the role of the orbitofrontal cortex and the striatum in the computation of goal values and prediction errors. *The Journal of Neuroscience*, *28*(22), 5623-30. doi:10.1523/JNEUROSCI.1309-08.2008

Hayes, T. R., Petrov, A. A., & Sederberg, P. B. (2011). A novel method for analyzing sequential eye movements reveals strategic influence on Raven ' s Advanced Progressive Matrices. *Journal of Vision*, *11*, 1-11. doi:10.1167/11.10.10.Introduction

Koller, D., & Parr, R. (1999). Computing Factored Value Functions for Policies in Structured {MDP}s. *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence* (pp. 1332-1339). The AAAI Press/The MIT Press.

Li, J., McClure, S. M., King-Casas, B., & Read Montague, P. (2006). Policy Adjustment in a Dynamic Economic Game. *PLoS ONE*, *1*(1), e103. Public Library of Science.

McClure, S. M., Berns, G. S., & Montague, P. R. (2003). Temporal prediction errors in a passive learning task activate human striatum. *Neuron*, *38*(2), 339-46.

Montague, P. R., Dayan, Peter, & Sejnowski, T. J. (1996). A Framework for Mesencephalic Predictive Hebbian Learning. *The Journal of Neuroscience*, *16*(5), 1936-1947.

Niv, Y., Edlund, J. a., Dayan, P., & O'Doherty, J. P. (2012). Neural Prediction Errors Reveal a Risk-Sensitive Reinforcement-Learning Process in the Human Brain. *Journal of Neuroscience*, *32*(2), 551-562. doi:10.1523/JNEUROSCI.5498-10.2012

O'Doherty, John P., Dayan, Peter, Friston, K., Critchley, H., & Dolan, R. J. (2003). Temporal difference models and reward-related learning in the human brain. *Neuron*, *38*(2), 329-37.

O'Doherty, John P., Dayan, Peter, Schultz, J., Deichmann, R., Friston, K., & Dolan, R. J. (2004). Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science*, *304*(5669), 452-4. doi:10.1126/science.1094285

Preuschoff, K., Bossaerts, P., & Quartz, S. R. (2006, August 3). Neural Differentiation of Expected Reward and Risk in Human Subcortical Structures. *Neuron*. Cell Press,.

Ribas-Fernandes, J. J. F., Solway, A., Diuk, C., McGuire, J. T., Barto, A. G., Niv, Yael, & Botvinick, M. M. (2011). A neural signature of hierarchical reinforcement learning. *Neuron*, *71*(2), 370-9. doi:10.1016/j.neuron.2011.05.042

Schultz, W., Dayan, Peter, & Montague, P. R. (1997). A Neural Substrate of Prediction and Reward. *Science*, *275*(March 1997), 1593-1599.

Schönberg, T., Daw, N. D., Joel, D., & O'Doherty, John P. (2007). Reinforcement Learning Signals in the Human Striatum Distinguish Learners from Nonlearners during Reward-Based Decision Making. *The Journal of Neuroscience*, *27*(47), 12860-12867. doi:10.1523/JNEUROSCI.2496-07.2007

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement Learning: An Introduction*. The MIT Press.

Sutton, R. S., Precup, D., & Singh, S. (1999). Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, *112*(1-2), 181-211. doi:10.1016/S0004-3702(99)00052-1

Footnotes

[i] It is worth noting that you could also make small updates to your casino-level knowledge after each game; this would imply two different prediction errors after *every* game, including at the end of the last game in the casino.

Divide and conquer: hierarchical reinforcement learning and task decomposition in humans

Carlos Diuk, Anna Schapiro, Natalia Cordova, Yael Niv and Matthew Botvinick

Department of Psychology and Princeton Neuroscience Institute
Princeton University
Princeton, NJ
{cdiuk,schapiro,ncordova,yael,matthewb}@princeton.edu

Abstract. The field of computational reinforcement learning (RL) has proved extremely useful in research on human and animal behavior and brain function. However, the simple forms of RL considered in most empirical research do not scale well, making their relevance to complex, real-world behavior unclear. In computational RL, one strategy for addressing the scaling problem is to introduce hierarchical structure, an approach that has intriguing parallels with human behavior. We have begun to investigate the potential relevance of hierarchical RL (HRL) to human and animal behavior and brain function. In the present chapter, we focus on one aspect of this work, which deals with the question of how action hierarchies are initially established. Work in HRL suggests that hierarchy learning is accomplished by identifying useful subgoal states, and that this might in turn be accomplished through a structural analysis of the given task domain. We review results from a set of behavioral experiments, in which we have investigated the relevance of these ideas to human learning and decision making.

1 Introduction

Many of the activities and tasks faced by humans and animals are hierarchical in nature: they involve tackling a set of nested subtasks, each of varying temporal extension. Problems like navigating involve devising high-level path plans, which are then broken down into smaller sub-planning problems, that can further be decomposed all the way down to the level of motor primitives. For instance, the task of commuting to work involves deciding whether to take a train, bus or drive, and based on that decision others must be made: taking a train will require navigating to the train station, driving might involve subtasks like filling up the gas tank or checking the state of traffic on the planned route. A hierarchical structure of nested tasks emerges, which will at some level share components like standing up, sitting down, walking and climbing stairs.

Work in cognitive and developmental psychology has recognized the hierarchical structure of behavior at least since the early 1950's, with the inception of the cognitive revolution. Prior to that watershed, the dominant schools of

thought had focused on understanding behavior as a simple chain of stimulus-response associations. Lashley (1951) rejected this idea in favor of understanding behavioral sequences as controlled through a central plan, rather than as simple reflex chains. Following up on this perspective, further pioneering work by Miller et al. (1960) and Schank and Abelson (1977) noted that naturalistic behavior displays a stratified or layered organization, comprising nested subroutines.

In subsequent years, the hierarchical structure of behavior has been taken for granted in psychology and neuroscience. Computational models have been proposed to account for how hierarchically structured procedures are represented and executed (Botvinick and Plaut, 2004; Cooper and Shallice, 2000; Schneider and Logan, 2006; Zacks et al., 2007), and how they are represented in the brain, in particular within the prefrontal cortex (Badre, 2008; Koechlin et al., 2003). An important idea, coming primarily out of developmental psychology, is that humans and other animals gradually expand their competence by building up a repertoire of reusable skills or subroutines, which can be flexibly assembled into increasingly powerful hierarchical programs of action (Fischer, 1980). The question of how this toolbox of skills is assembled represents one of the toughest questions attaching to hierarchical behavior.

In recent work, we have adopted a novel perspective on the cognitive and neural mechanisms underlying hierarchical behavior, leveraging tools from machine learning research. In particular, we have examined the potential relevance to human behavior and brain function of hierarchical reinforcement learning (HRL), a computational framework that extends reinforcement learning mechanisms into hierarchical domains. A number of intriguing parallels exist between HRL and findings from human and animal neuroscience, which encourage the idea that HRL may provide a useful framework for understanding the biological basis of hierarchical behavior. In the following section, we briefly review the essentials of HRL and summarize some of the potential neuroscientific parallels. However, our main purpose in the present chapter is more focused. One appealing aspect of HRL is that it provides a context within which to consider the "toolbox" question, the question of how useful skills or subroutines are initially discovered or constructed. Following our brief introductory survey, we describe a set of behavioral experiments in which we have leveraged ideas from HRL to tackle this question.

2 Hierarchical Reinforcement Learning

Computational reinforcement learning (RL) has emerged as a key framework for modeling and understanding decision-making in humans and animals. In part, this is due to the fact that RL provides a normative computational model of behavior accounting for a host of previous experimental results in classical and instrumental conditioning. But most importantly, its impact has been felt through the discovery of parallels between elements of RL and aspects of neural function. The most critical parallel pertains to midbrain dopaminergic function, which has been proposed to transmit signals comparable to the reward-prediction errors

that lie at the heart of RL (Barto, 1995; Montague et al., 1996; Schultz et al., 1997). However, other broader parallels have also been proposed, in particular with so-called actor-critic RL architectures, which have inspired new interpretations of functional divisions of labor within the basal ganglia and cerebral cortex (Joel et al., 2002). Our research asks whether these connections between RL and neurobiology might extend to the setting of hierarchical behavior. Based on the success of standard RL as a framework for understanding the neural mechanisms underlying simple decision making, we hypothesize that HRL may hold similar promise as a framework for understanding the neural basis of hierarchical action.

Computational HRL was born, in part, out of the attempt to tackle the problem of scaling in RL. As researchers in the field recognized early on, one of the problems of basic RL methods is that they cannot cope well with large domains, that is, problems that require learning about large numbers of world states or large sets of possible actions. To make matters worse, RL suffers from what is known as the *curse of dimensionality*, an exponential explosion in the number of states as we increase the number of state variables, or features of the problem, that we want to consider. The result is that any task that requires keeping tabs on more than a handful of variables soon becomes intractable for standard RL algorithms.

A number of computational approaches have been proposed to address the scaling issue. One of them is to reduce the size of the problem at hand by treating subsets of states as behaviorally equivalent, known as *state abstraction*. Consider for example that you are walking to the train station, on your way to work. For this task, whether the shops along the way are open or closed is irrelevant, so two states that only differ in the status of a store can be grouped together. On the other hand, if later on you are navigating the same streets with the goal of buying coffee, a different set of variables becomes relevant, and states should be abstracted differently. For different state abstraction methods and aggregation criteria see Li et al. (2006).

Another approach to addressing the scaling problem – the one taken in HRL – is based on *temporal abstraction* (Barto and Mahadevan, 2003; Dietterich, 2000; Parr and Russell, 1998; Sutton et al., 1999). The general idea is to expand the standard RL framework to include temporally-extended macro-actions, grouping together sets of simpler actions to form more complex, higher-level routines. Following the example mentioned earlier, the skill of *getting to work* can be thought of as a representation for a set of lower-level sequences like walking to the train station, taking the train and walking from the station to work. Moreover, the same *get to work* skill can encompass more than one set. For example, this skill might not only consist of a set of actions involving the train, but also a different set that consists of actions like walking to the car, starting it, driving to work, etc. These multiple representations, abstracted away into the skill of *getting to work*, enable learning and reasoning at a coarser, more tractable granularity.

One particularly influential implementation of HRL, the *options* framework, was proposed by Sutton et al. (1999). The options framework supplements the

set of single-step, primitive actions from standard RL with a set of temporally-extended "options." An option is, in a sense, a temporary sub-policy, a mapping from states to actions that does not have the goal of solving the complete problem at hand, but rather some sub-task that is, ideally, a step towards a larger goal. In this formalism, an option is defined by an initiation set, indicating the set of states from which the option can be selected; a termination function, which specifies the set of states that trigger termination of the option; and an option-specific policy (a mapping from states to actions that is in effect while the option is active).

Importantly, in the options framework as in other versions of HRL (Dietterich, 2000; Parr and Russell, 1998), option-specific policies can map states not only into primitive actions but also into other options, allowing hierarchies of options to be assembled. In the previous example, it is clear that walking to a train station or to the car are not "primitive" actions, but compound, temporally extended behaviors that involve numerous more basic skills, and can be achieved in a multiplicity of ways. In an HRL setting, an option for getting to work would call other options for walking to the train station or the car, these would call further options guiding the action of walking, and so forth down to elementary motor commands.

3 Potential Neural Correlates

We see two reasons for considering the potential relevance of HRL to understanding behavior and brain function in humans and other animals. First, if the brain does indeed implement learning mechanisms related to those found in RL, then the RL scaling problem must pertain in neuroscience just as it does in machine learning, raising the question of how RL mechanisms in the brain cope with large-scale tasks. As a computational technique for easing the scaling problem, HRL may furnish clues concerning the brain's ability to select adaptive behaviors in such settings. The second motivation for considering HRL from a neuroscientific perspective is, of course, the pervasively hierarchical structure of human behavior. HRL presents the possible opportunity to extend our understanding of neural mechanisms for RL so as to engage the issue of hierarchy, significantly widening the scope of current theories.

As a first step toward evaluating the potential neural relevance of HRL, Botvinick et al. (2009) derived a set of predictions from the framework, evaluating the extent to which current scientific knowledge accorded with each of its elements. This work leveraged the existence of proposed parallels between elements of the actor-critic architecture for RL (see Sutton and Barto, 1998) and specific brain structures. Botvinick et al. considered what additions or alterations would be required in order to extend the actor-critic architecture for HRL. It turns out that only a handful of modifications are needed, and each of these appears to resonate with established neuroscientific findings.

A key parallel pointed out by Botvinick et al. (2009) relates to the computational requirement, within HRL, of maintaining a representation of the currently

selected option. This function seems very closely related to functions commonly ascribed to the dorsolateral prefrontal cortex (DLPFC), and other frontal areas including pre-supplementary motor area (pre-SMA). The DLPFC has been suggested to house representations that guide temporally integrated, goal-directed behavior (Fuster, 1997), and recent work has refined this idea by demonstrating that DLPFC neurons play a role in representing task sets: a single pattern of DLPFC activation represents an entire mapping from stimuli to responses (that is, a policy; see Miller and Cohen, 2001). Moreover, neurons in several frontal areas (DLPFC, pre-SMA and SMA) have been shown to code for particular sequences of low-level actions, just like options do in HRL. Evidence also shows that areas in frontal cortex represent action at multiple, nested levels of temporal structure (see Badre, 2008; Koechlin et al., 2003), akin to the way HRL representations organize tasks into hierarchies, with policies for one option calling other, lower-level options.

The role of options in HRL is to impose an option-specific policy. In translations of RL into neuroscience, policy representations have been proposed to reside at least partially within the dorsolateral striatum. From the point of view of the HRL hypothesis, it is suggestive that DLPFC, SMA and pre-SMA areas all project heavily into this structure, potentially allowing modulation of policy representations by representations of subtask context. Botvinick et al. (2009) review neurophysiological findings consistent with this idea.

Another computational requirement of HRL is to maintain option-specific value functions. As discussed in Botvinick et al. (2009), this is needed because the value of a state relative to the goals of an option or subroutine may differ from the value of that state relative to top-level goals (i.e., primary reward); option-specific value functions are thus critical for driving the learning of subroutine policies. In work drawing parallels between standard RL and neural structures, an area often linked with state or state-action value representation is the ventral striatum. If HRL mechanisms are relevant, then we might expect to find a neural structure that connects to ventral striatum while at the same time receiving inputs from areas of frontal cortex that carry option representations. An area that meets this criterion is the orbitofrontal cortex (OFC), connecting heavily with both ventral striatum and DLPFC. As reviewed by Botvinick et al. (2009), research suggests that representations of reward in OFC can be sensitive to shifts in response strategy or task set (O'Doherty et al., 2003; Schoenbaum et al., 1999), linking precisely with the idea that OFC might represent option-specific state values. The OFC also appears to sustain reward-predictive activity over relatively extended periods (Schultz et al., 2000), a function necessary in HRL to support the calculation of reward-prediction errors when options terminate.

As detailed in Botvinick et al. (2009), neural HRL would also impose specific functional requirements on reward-prediction errors, widely believed to be signaled in the brain by phasic fluctuations in dopamine release. Whereas in ordinary RL prediction errors signal whether the selection of single actions turns out better or worse than expected (see Sutton and Barto, 1998), under HRL the scope of the prediction error expands to embrace the intervals spanned by op-

tions. This resonates with a theoretical analysis of dopamine signaling by Daw et al. (2003), interpreting dopamine function in computational (semi-Markov) terms that also underlie the options framework. Another neural prediction from HRL is that reward prediction errors should occur not only in association with top-level goals (marked by primary reward), but also in connection with *subgoals*, and their associated pseudo-reward. In this case, previous research provides little to go on. However, Ribas-Fernandes et al. (2010) used EEG and fMRI to assay for subgoal-linked reward prediction errors and found activations consistent with these in multiple structures including anterior cingulate cortex, insula, habenula, amygdala, and ventral striatum.

Taken together, available neural data encourage the idea that HRL may be relevant to understanding the neural substrates of hierarchical behavior in humans and animals. Even if this turns out to be true, however, there are limits on what present-day HRL research can tell us about brain function, given that computational HRL is associated with its own open questions. Perhaps foremost among these is the problem foreshadowed earlier: how an agent may initially build up a repertoire of useful subroutines (options) from which hierarchical action programs may be composed. This question, which in HRL research has sometimes been referred to as the "option discovery problem," is clearly of equal importance within psychology and neuroscience, and we now turn to work in which we have begun to address it.

4 The option discovery problem: Identifying useful subgoals

In the field of computational HRL, research has focused on the problem of how temporally-extended actions can be incorporated into the standard RL formalism. Some success has been achieved in showing how skills that are provided to the learner as input, or have somehow been previously acquired, can be exploited in order to learn to solve new problems faster (Dietterich, 2000; Sutton et al., 1999). However, less work has been done, and less success has been achieved, on the very difficult question of where skill representations come from. How does a learner decide, while performing a task, what components of it are worth incorporating into a collection of skills for future use?

In computational work, option discovery has often been understood to involve the heuristic identification of useful *subgoal* states. Once a useful subgoal is identified, the learner can then build a strategy to achieve it, turning this strategy or policy into a reusable skill. Note that these subgoal states are not necessarily extrinsically rewarding, that is, the learner might not receive any reward for reaching them. A key assumption of HRL is that the agent is "intrinsically" motivated to reach an options subgoal, once the option gains control of behavior. Instead, in HRL attaining the subgoal yields a special reward signal, referred to as *pseudo-reward*, which serves to sculpt the options policy. However, for this machinery to come into play, this pseudo-reward function must be

established, and therefore the question persists: How are useful subgoal states initially identified?

A number of possible answers to this question can be drawn from both the computational literature, and from psychology and neuroscience. One class of proposals has thought of options as genetically specified, shaped by natural selection across generations (Elfwing et al., 2007). Basic motor behavior, for example, has often been characterized as building upon simple, innate components (Bruner, 1975). In a few cases, extended action sequences, such as grooming in rodents, have also been thought of in the animal behavior literature as genetically specified (Aldridge and Berridge, 1998). While a role for evolutionary programming seems inevitable, it clearly cannot be the whole story, since both humans and animals obviously discover and incorporate useful behavioral subroutines through learning (Conway and Christiansen, 2001; Fischer, 1980).

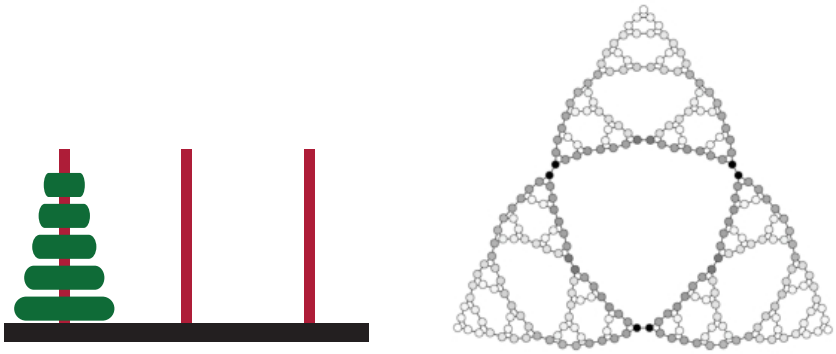
Another approach to explaining subgoal discovery leverages the notion of intrinsic motivation. The idea here is that certain events or stimuli are inherently interesting to the behaving animal or human. These can be stimuli that display salient perceptual properties or that challenge expectations, eliciting curiosity Schmidhuber (1991a,b). In an HRL context such states are proposed to be adopted as subgoals, at least temporarily until their properties are properly learned, triggering the construction of associated skills or options (see Barto et al., 2004).

The intrinsic motivation perspective provides a compelling account of option discovery. However, without greater specification, it leaves open the question of *which* properties make particular states intrinsically motivating or interesting to the agent. Potential answers to this question are explored in several chapters in the present volume. In order to set the scene for our own research in this area, we can consider two general approaches, one based on frequency and the other on problem structure.

Frequency-based methods are based on observed trajectories (that is, sequences of actions that are performed to solve a task). These methods are based on the idea that an animal or human that has experienced a series of interrelated problems, or has had repeated exposure to a problem, is able to extract either subsequences or sub-goal states based on their frequent occurrence in trajectories that lead to reward. For example, consider a delivery person distributing packages inside a building. After repeated deliveries, this person might construct some pre-defined ways of traversing certain floors. Furthermore, he might realize that many trajectories involve taking the elevator. He would thus identify reaching the elevator as a useful sub-goal, and construct paths that lead from different offices in a floor to the closest elevator, adding to his repertoire of actions what we could call the *go to elevator* option. Proposals based on this idea can be found in the work of Mcgovern and Barto (2001); Picket and Barto (2002); Thrun and Schwartz (1995); Yamada and Tsuji (1989).

To introduce structure-based methods it is useful to consider why, in the aforementioned delivery example, the elevator state emerged as special. In this scenario, the elevator state occurs frequently because the elevator is a sort of

bottleneck: to reach any location on one floor from another floor, one must pass through the elevator. The elevator is thus a location that gives access to an unusually diverse set of other locations. A more formal way of capturing this property can be drawn from graph theory. If we envision the various locations (say, cubicles and offices in our couriers building) as nodes in a graph, with edges connecting immediately adjacent locations, then the elevator location would stand out as a node with high graph *centrality* (see Opsahl et al., 2010). A particular way of quantifying centrality is via a measure called *betweenness*, which counts the number of shortest paths within the graph that pass through an index node. An illustration, from Şimşek (2008), is shown in Figure 1.



(a) (b)

Fig. 1. (a) One state of the Tower of Hanoi problem. Disks are moved one at a time between posts, with the restriction that a disk may not be placed on top of a smaller disk. An initial state and goal state define each specific problem. (b) Representation of the Tower of Hanoi problem as a graph. Nodes correspond to states (disk configurations). Shades of gray indicate betweenness. Source: Şimşek (2008).

Şimşek (2008) and Şimşek and Barto (2009) proposed that option discovery might be fruitfully accomplished by identifying states at local maxima of graph betweenness (for related ideas, see also Şimşek et al. (2005); Hengst (2002); Jonsson and Barto (2006); Menache et al. (2002). They presented simulations showing that an HRL agent designed to select subgoals (and corresponding options) in this way, was capable of solving complex problems, such as the Tower of Hanoi problem in Figure 1(a), significantly faster than a non-hierarchical RL agent.

As part of our research exploring the potential relevance of HRL to neural computation, we evaluated whether these proposals for subgoal discovery might relate to procedures used by human learners. The research we have completed so

far focuses on the identification of bottleneck states, as laid out by Şimşek and Barto (2009). In what follows, we summarize the results of three experiments, which together support the idea that the notion of bottleneck identification may be useful in understanding human subtask learning.

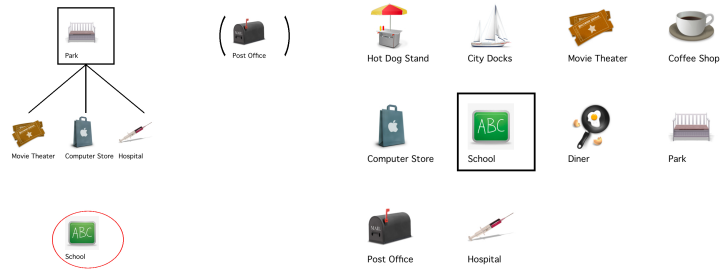
5 Experiments 1 & 2: Humans Identify and Exploit Bottleneck States

In a first experiment we investigated whether humans can identify bottleneck states, when doing so allows them to optimize their performance. We summarize the experiment and its results here; full details are presented in Cordova et al. (2010).

Participants were asked to navigate through a small town, making an extended series of deliveries between landmarks (e.g., school, post office, coffee shop). A new start location and goal location were randomly selected at the beginning of each trial (delivery). Participants were told that they would be paid for each delivery, but that the amount would depend on how many steps they took to reach their goal: each step would subtract a fixed amount from the full pay'. The graphical interface, illustrated in Figure 2(a), indicated the participants present location, the goal location, and the set of landmarks immediately adjacent to it. Navigation was accomplished by selecting among the latter. Also shown was a "bus stop" location to which the participant could travel from any location using one step. After some experience with the "town", the participant was allowed to choose a new bus-stop location after every five deliveries. Any landmark within the town could be chosen for the bus-stop location. At any time during a delivery, the participant could elect to "jump" to the bus stop, potentially saving costly steps toward the goal.

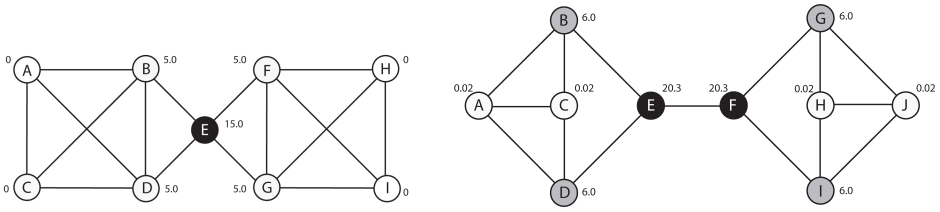
Underlying the adjacency relations among landmarks in the town was the graph shown in Figure 3(a). Each node corresponds to a landmark, and each edge to an adjacency relation. The graph contains an obvious bottleneck location, which has high graph betweenness. This location represents the best choice for the bus-stop location; given the definition of betweenness, this location lies on the largest number of shortest paths within the graph, and therefore offers the best chance of saving the participant steps toward a delivery to a yet-unknown destination.

Note that participants were never actually shown the graph in Figure 3(a), or any other sort of birds-eye view of the town. The display only provided information about local adjacencies. Nevertheless, we hypothesized that, with accumulating experience, participants would identify the bottleneck location and exploit it by selecting it as a bus-stop location. Figure 4 summarizes the results of the experiment. Panel *a* shows that, over the course of the experiment, participants increasingly picked out the shortest path from start to goal. This simply provides evidence that participants learned something about the layout of the town as they went along. More important are the data in panel *b*, which show the number of blocks (out of a total of 16) in which each participant chose to



(a) First version. (b) Second version

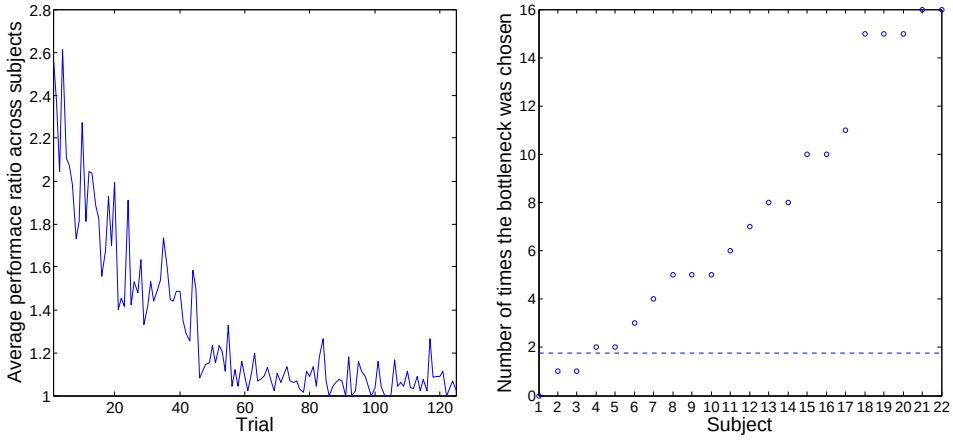
Fig. 2. Interface of experiments 1 & 2. (a) At the top, the current location (*Park*) is identified along with its three adjacent locations. Circled at the bottom is the target destination (*School*), and on the upper right corner is the bus-stop location (*Post office*), reachable in one step from any other location. (b) The square identifies the current location (*School*), and participants must click on its three neighbors.



(a) (b)

Fig. 3. Graphs underlying the maps of the cities for the first version of the experiment (a) and the second one (b). Node labels identify the betweenness of each node.

place the bus stop at the bottleneck location. Although there was some variability across participants, the data clearly confirm a general capacity to detect and exploit the presence of a bottleneck.



(a) (b)

Fig. 4. (a) Average performance ratio, over all participants, as a function of trials of experience with the "town". The value on the y-axis in the figure represents the ratio of steps taken to the minimum number of steps, taking into account the optimal bus stop location. A ratio of 1 indicates optimal performance, i.e., choice of the shortest path from start to goal, assuming an optimal (bottleneck) choice of bus stop. (b) Number of times, out of 16 five-trial blocks, that the bottleneck state was chosen for the bus-stop location. Participants in the x-axis are sorted by performance. The dashed line indicates the expected performance if participants chose bus-stop locations randomly.

The results of this experiment do not, however, allow us to make conclusions about *how* participants identified the bottleneck location. In particular, while we were interested in the possibility that they leveraged structural or topological knowledge, it is possible that participants instead used simple frequency information. Over the course of multiple shortest-path deliveries, the bottleneck location would be expected to occur frequently, compared with other locations. In order to rule out frequency as the full explanation for our initial findings, we repeated Experiment 1, but with a twist. In this revised version (Experiment 2), participants learned about adjacency relations, but did not ever traverse the town before choosing a bus-stop location. This follow-up experiment was also intended to address a second limitation of the first experiment. Note that in the graph used in the first experiment, not all vertices had the same degree (i.e., the same number of immediate neighbors). While vertices on the outskirts of the

city had three neighbors, the bottleneck vertex and those adjacent to it had four. In principle, this might have made the bottleneck salient, providing a different explanation for its selection.

Experiment 2 removed the confound between centrality and frequency, and used a graph in which all vertices had the same degree (Figure 3(b)). Figure 2(b) illustrates the graphical interface for the task. On each trial, an index location was highlighted, and participants were asked to indicate its three immediate neighbors, receiving feedback concerning the accuracy of their choices. After approximately twenty minutes on this training task, participants were told they would have to make a delivery between two undisclosed locations, under the same shortest-path conditions as in Experiment 1. Prior to receiving the delivery assignment, participants were asked to choose a location for the bus stop. After they had chosen a location, their knowledge of the underlying topology of the town was tested by asking them to draw a map, indicating adjacency relations between landmarks. Of forty participants tested, 23 drew an accurate map of the town, and of these 23, 18 (78%) chose one of the bottleneck locations as the bus stop location, a result far above the chance level of 20%.

Together, the results of Experiments 1 and 2 provide support for the idea that humans can identify and exploit bottleneck states in a novel domain, based on an internal model of the domains structure. Taken on their own, however, they leave open a second question. The computational proposal from HRL was that bottleneck locations provide the anchor for temporal abstractions, representations that treat temporally extended behaviors as a unit. The experiments just reported do not speak to this aspect of the theoretical proposal. However, we can glean some pertinent evidence from a third experiment.

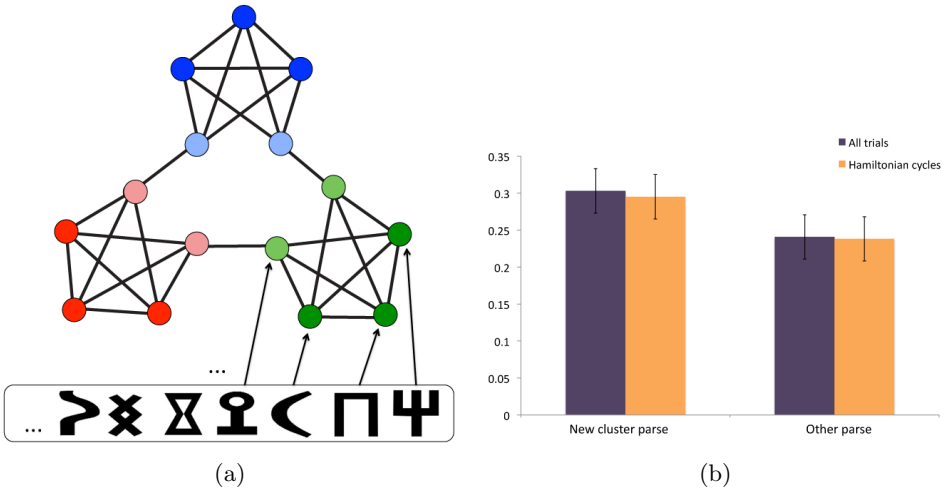
6 Experiment 3: Bottleneck States and Temporal Abstraction

Our approach in Experiment 3 was based on previous work using event parsing. A standard experimental paradigm in cognitive psychology involves showing an action sequence, and asking participants to parse it by pressing a key when they feel that one subsequence or subtask has ended and a new one has begun (Zacks et al., 2007). Consistent with earlier work, we assumed that such parsing responses mark the boundaries of temporally abstract events, i.e., subsequences that the participant views, on some level, as a unit. Based on this assumption, we predicted that if participants were exposed to event sequences that involved bottlenecks, participants would parse those sequences at moments in which a bottleneck was traversed. Details of our experiment are reported in Schapiro et al. (2010); we summarize the work here.

Participants were exposed to a sequence of images presented, one per second, over a period of 35 minutes. Every image presented came from a set of fifteen, as shown in Figure 5(a). During this exposure period, participants were asked to judge whether each image was presented in a canonical orientation, or rotated. The task did not require them to attend to the sequential order of images at

all. However, unbeknownst to the participants, that order was highly structured. Specifically, the sequence was generated by a random walk through the graph shown in Figure 5(a). Each image was assigned to a vertex, and when that vertex occurred in the random walk, the associated image was presented. As is obvious from the figure, the graph contains a subset of bottleneck vertices with high betweenness, namely the vertices that link the three star-shaped clusters.

After performing the orientation judgment task, the sequence of images continued, but participants were asked to perform the standard parsing task, pressing a key when natural breakpoints occurred, i.e., when one subsequence ended and a new one began. No instruction other than this was given.



(a) (b)

Fig. 5. (a) Underlying graph of the task. Each node in the graph is linked to a character used in the sequence. (b) Proportion of times participants parsed sequence at cluster-changing points, as opposed to other characters in the sequence.

Results indicated that participants were significantly more likely to parse at moments where the sequence moved from one star-shaped cluster into another ($p < 0.05$), points corresponding to the traversal of high-betweenness vertices. This result held even when the analysis was limited to Hamiltonian cycles through the graph (traversals of the graph without item repetitions), showing that parsing decisions were not based entirely on item recency judgments or simple effects of priming.

The relation of this experiment to HRL-like action selection is necessarily indirect, given that the task involved observation rather than production of sequences. However, the results are in line with the idea that bottleneck states are not only spontaneously identified by humans, but that bottlenecks provide

a basis for the formation of temporally abstract event representations. This is consist with the proposal that bottleneck states provide anchors for the construction of temporally abstract action representations, i.e., options, although further experimentation will be needed to validate this inference.

7 Discussion

The development of the field of computational RL, together with the discovery of its neural implications, has proven extremely useful in the study of human and animal behavior and brain function. A known limitation of standard RL, however, is its poor scaling to large, real-world problems. Given this limitation, it is unreasonable to expect basic RL principles to account for human learning and decision making in their full complexity. However, the possibility arises of looking at measures proposed by the computational community to deal with the scaling problem, evaluating their possible relevance to the biological case. We reported work that takes this approach, examining one aspect of complex behavior, namely its hierarchical structure. In the work we have reported, the aim was to leverage existing work in HRL, a sub-field developed precisely for tackling the scalability problem, to shed light on how humans might learn to master hierarchically-structured tasks. Our agenda was further reinforced by evidence of potential neural correlates that map nicely with existing HRL frameworks.

One aspect of hierarchical learning, which has provided an important focus for our work, involves the challenge of discovering useful subtask decompositions. On the computational front, this problem has suggested a form of intrinsic motivation, which leads learning agents to identify problem states as sub-goals, constructing the necessary skills to achieve them. The work we have reviewed tested the relevance of this idea to human learning and decision making. In particular, we explored one approach to this problem, based on structural task analysis. We presented three experiments whose results are consistent with the idea that humans are able to learn the topological structure underlying a problem domain, to detect states associated with high centrality (in the graph-theoretic sense), and to adopt them as useful subgoals and as anchors for temporally abstract event representations.

Overall, the work we have reviewed, together with convergent evidence available from previous studies, suggests that HRL may provide a useful set of tools for further investigating the computational and neural basis of hierarchically structured behavior. In this sense, HRL may play the same catalytic role, in the context of hierarchical behavior, that ordinary RL has so fruitfully played in the study of performance in simpler tasks.

Bibliography

Aldridge, J. W. and Berridge, K. C. (1998). Coding of serial order by neostriatal neurons: a "natural action" approach to movement sequence. *Journal of Neuroscience*, 18(7):2777–87.

Badre, D. (2008). Cognitive control, hierarchy, and the rostro-caudal organization of the frontal lobes. *Trends In Cognitive Sciences*, 12(5):193–200.

Barto, A. G. (1995). *Adaptive Critics and the Basal Ganglia*, pages 215–232. Number 1994. MIT Press, Cambridge, MA.

Barto, A. G. and Mahadevan, S. (2003). Recent Advances in Hierarchical Reinforcement Learning. *Discrete Event Dynamic Systems*, 13(4):341–379.

Barto, A. G., Singh, S., and Chentanez, N. (2004). Intrinsically Motivated Reinforcement Learning. *Advances in Neural Information Processing Systems 17 (NIPS)*.

Botvinick, M. and Plaut, D. C. (2004). Doing without schema hierarchies: a recurrent connectionist approach to normal and impaired routine sequential action. *Psychological review*, 111(2):395–429.

Botvinick, M. M., Niv, Y., and Barto, A. C. (2009). Hierarchically organized behavior and its neural foundations: a reinforcement learning perspective. *Cognition*, 113(3):262–80.

Bruner, J. (1975). Organization of early skilled action. *Child Development*, 44:1–11.

Conway, C. M. and Christiansen, M. H. (2001). Sequential learning in non-human primates. *Trends in Cognitive Sciences*, 5(12):539–546.

Cooper, R. and Shallice, T. (2000). Contention scheduling and the control of routine activities. *Cognitive neuropsychology*, 17(4):297–338.

Cordova, N., Diuk, C., Niv, Y., and Botvinick, M. (2010). Discovering hierarchical task structure. *In preparation*.

Daw, N. D., Courville, A. C., and Touretzky, D. S. (2003). Timing and partial observability in the dopamine system. In *Advances in Neural Information Processing Systems (NIPS)*, Cambridge, MA. MIT Press.

Dietterich, T. G. (2000). Hierarchical Reinforcement Learning with the MAXQ Value Function Decomposition. *Journal of Artificial Intelligence Research*, 13:227–303.

Elfwing, S., Uchibe, E., Doya, K., and Christensen, H. I. (2007). Evolutionary development of hierarchical learning structures. *IEEE transactions on evolutionary computation*, 11(2):249–264.

Fischer, K. W. (1980). A Theory of Cognitive Development: The Control and Construction of Hierarchies of Skills. *Psychological Review*.

Fuster, J. M. (1997). *he Prefrontal Cortex: Anatomy, Physiology, and Neuropsychology of the Frontal Lobe*. Lippincott-Raven, Philadelphia, PA, third edition.

Hengst, B. (2002). Discovering Hierarchy in Reinforcement Learning with HEXQ. *Proceedings of the 19th International Conference on Machine Learning*.

Joel, D., Niv, Y., and Ruppin, E. (2002). Actor critic models of the basal ganglia : new anatomical and computational perspectives. *Neural Networks*, 15:535–547.

Jonsson, A. and Barto, A. (2006). Causal Graph Based Decomposition of Factored MDPs. *J. Mach. Learn. Res.*, 7:2259–2301.

Koechlin, E., Ody, C., and Kouneiher, F. (2003). The architecture of cognitive control in the human prefrontal cortex. *Science (New York, N.Y.)*, 302(5648):1181–5.

Lashley, K. S. (1951). *The problem of serial order in behavior*. Wiley, New York.

Li, L., Walsh, T. J., and Littman, M. L. (2006). Towards a Unified Theory of State Abstraction for MDPs. In *Proceedings of the Ninth International Symposium on Artificial Intelligence and Mathematics (AMAI-06)*.

Mcgovern, A. and Barto, A. G. (2001). Automatic Discovery of Subgoals in Reinforcement Learning using Diverse Density. *Proc of the 18th International Conference on Machine Learning*.

Menache, I., Mannor, S., and Shimkin, N. (2002). Q-cutdynamic discovery of sub-goals in reinforcement learning. *European Conference on Machine Learning (ECML 2002)*, pages 295–306.

Miller, E. K. and Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual review of neuroscience*, 24:167–202.

Miller, G. A., Galanter, E., and Pribram, K. H. (1960). *Plans and the structure of behavior*. Adams-Bannister-Cox, New York.

Montague, P. R., Dayan, P., and Sejnowski, T. J. (1996). A Framework for Mesencephalic Predictive Hebbian Learning. *Journal of Neuroscience*, 16(5):1936–1947.

O'Doherty, J., Critchley, H., Deichmann, R., and Dolan, R. J. (2003). Dissociating valence of outcome from behavioral control in human orbital and ventral prefrontal cortices. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 23(21):7931–9.

Opsahl, T., Agneessens, F., and Skvoretz, J. (2010). Node centrality in weighted networks: Generalizing degree and shortest paths. *Social Networks*, 32:s.

Parr, R. and Russell, S. J. (1998). Reinforcement learning with hierarchies of machines. *Advances in neural information processing systems*.

Picket, M. and Barto, A. (2002). PolicyBlocks: An Algorithm for Creating Useful Macro-Actions in Reinforcement Learning. In *Proceedings of the 19th International Conference on Machine Learning*.

Ribas-Fernandes, J. J. F., Niv, Y., and Botvinick, M. M. (2010). Neural correlates of Hierarchical Reinforcement Learning. *Submitted*.

Schank, R. C. and Abelson, R. P. (1977). *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures.* Lawrence Erlbaum, Hillsdale, NJ.

Schapiro, A., Rogers, T., Cordova, N., Turk-Browne, N., and Botvinick, M. M. (2010). The structure of event representations: behavioral, imaging, and computational investigations. *In preparation*.

Schmidhuber, J. (1991a). A possibility for implementing curiosity and boredom in model-building neural controllers. *Proceedings of the International Conference on Simulation of Adaptive Behavior: from Animals to Animats*, pages 222–227.

Schmidhuber, J. (1991b). Curious model-building control systems. *Proceedings of the International Conference on Neural Networks*, 2:1458–1463.

Schneider, D. W. and Logan, G. D. (2006). Hierarchical control of cognitive processes: switching tasks in sequences. *Journal of experimental psychology. General*, 135(4):623–40.

Schoenbaum, G., Chiba, a. a., and Gallagher, M. (1999). Neural encoding in orbitofrontal cortex and basolateral amygdala during olfactory discrimination learning. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 19(5):1876–84.

Schultz, W., Dayan, P., and Montague, P. R. (1997). A Neural Substrate of Prediction and Reward. *Science*, 275(March 1997):1593–1599.

Schultz, W., Tremblay, L., and Hollerman, J. R. (2000). Reward processing in primate orbitofrontal cortex and basal ganglia. *Cerebral cortex*, 10(3):272–84.

Şimşek, O. (2008). *Behavioral building blocks for autonomous agents: description, identification, and learning*. PhD thesis, University of Massachussetts, Amherst.

Şimşek, O. and Barto, A. G. (2009). Skill Characterization Based on Betweenness. In Koller, D., Schuurmans, D., Bengio, Y., and Bottou, L., editors, *Advances in Neural Information Processing Systems 21*, pages 1497–1504.

Şimşek, O., Wolfe, A. P., and Barto, A. G. (2005). Identifying useful subgoals in reinforcement learning by local graph partitioning. In *Proceedings of the Twenty-Second International Conference on Machine Learning*.

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

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

Thrun, S. and Schwartz, A. (1995). Finding Structure in Reinforcement Learning. In Tesauro, G., Touretzky, D., and Leen, T., editors, *Advances in Neural Information Processing Systems (NIPS) 7*, Cambridge, MA. MIT Press.

Yamada, S. and Tsuji, S. (1989). Selective learning of macro-operators with perfect causality. In *Proceedings of the 11th international joint conference on Artificial intelligence - Volume 1*, pages 603–608, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Zacks, J. M., Speer, N. K., Swallow, K. M., Braver, T. S., and Reynolds, J. R. (2007). Event perception: a mind-brain perspective. *Psychological bulletin*, 133(2):273–93.