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

Carlos Diuk, Anna Schapiro, Natalia Cordova, Jose Ribas-Fernandes, Yael Niv and Matthew Botvinick

Department of Psychology and Princeton Neuroscience Institute
Princeton University
Princeton, NJ
{cdiuk,schapiro,ncordova,jf3,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 first review two results that show the existence of neural correlates to key predictions from HRL. Then, 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 and neuroimaging 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; Haruno and Kawato, 2006; Ito and Doya, 2011; 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. We then present results suggesting neural correlates to two key predictions arising from computational HRL models. Next, we focus on a deep and open question: how is hierarchical structure established? What constitutes a "good" task decomposition? 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 and neuroimaging 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 in-

strumental 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 the 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; Dayan and Hinton, 1993; 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 al-

terations 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 options. 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.

Two key neural predictions arise from HRL. First, in order to sustain learning of option-specific value functions at various levels of a hierarchical task decomposition, multiple prediction error signals are required, sometimes occurring concurrently. Second, HRL predicts that reward prediction errors should occur not only in association with top-level goals (marked by primary reward), but also in connection with *subgoals*. In both cases, previous research provides little to go on. In the next two sub-sections we present work indicating the presence of neural correlates to the two key predictions from HRL. First, we summarize results from an fMRI experiment (Diuk et al., 2012b) which revealed striatal activity correlating with two simultaneous prediction error signals, corresponding to two levels of a hierarchical gambling task. Next, we review work by Ribas-Fernandes et al. (2011) in which we 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 and amygdala.

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 will dedicate the rest of the chapter to work in which we have begun to address it.

3.1 Two simultaneous, but separable, reward prediction errors

Under the HRL options framework, a situation can arise in which the outcome of an action elicits learning at multiple hierarchical levels at the same time. For example, the execution of a primitive action a that leads to a sub-goal state enables learning both about the one-step transition produced by the action, as well as the temporally-extended subtask that ended with the attainment of the sub-goal. This situation prescribes the presence of two distinct reward prediction errors.

If the brain implements an HRL mechanism, we should be able to measure activity correlating with at least two prediction errors at the same time. In order

to test this key HRL prediction, we designed a two-level gambling task (Diuk et al., 2012b), which constitutes a hierarchical extension of the classic bandit task used in previous RL research. The task is summarized in Figure 1. 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.

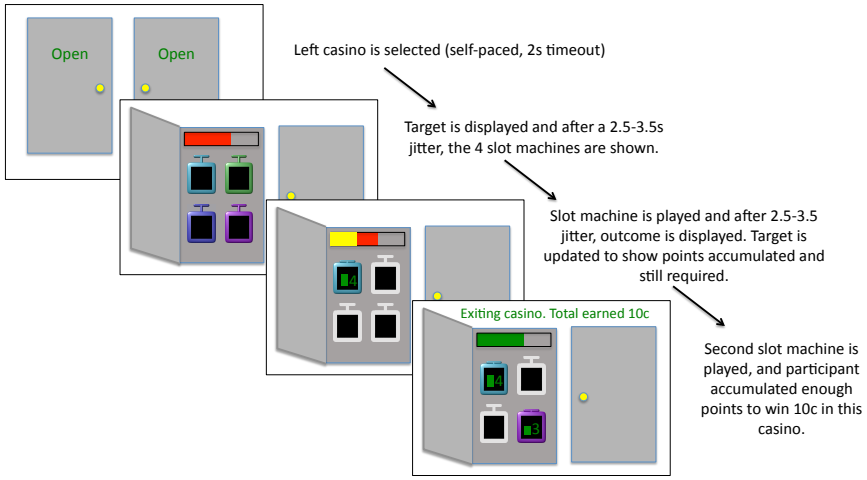


Fig. 1. Sample trial (from (Diuk et al., 2012b): The participant chooses to play in the left casino, the door opens and displays a target number of points (indicated by red bar). After a few seconds, the four slot machines appear. The participant plays upper-left slot and after a few seconds, the points obtained in that machine are shown inside the machine (as a green bar plus a Roman numeral). 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. The participant plays the bottom right slot machine and obtains sufficient points to win the casino (10 cents). The target bar turns green and a message appears indicating the casino win.

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 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.

We asked 28 participants to play the Casino Task for 120 trials each while undergoing functional Magnetic Resonance Imaging (fMRI) (Diuk et al., 2012b). We modeled the participants' learning under the options framework, where playing the left or the right casino constituted two temporally-extended options and each options' policy consisted of choosing which slot machines to play. We verified that the HRL model best fit the participants behavior when compared to some otherwise plausible alternative models, and used the prediction errors generated by this model as regressors to correlate against the registered brain activity. We analyzed in particular the activity in ventral striatum and found it correlated with all three regressors generated by the model, corresponding to prediction errors for the first slot machine ($p < 0.004$), second slot machine ($p < 0.02$) and the casino as a whole ($p < 5.5 \times 10^{-5}$). Note that prediction errors for the second slot machine and the casino occurred simultaneously.

These results have two major implications: The first is that the human brain can calculate prediction errors that temporally span over several states and actions, as is required in HRL (Botvinick et al., 2009). The second implication is that more than one prediction error signal may be calculated and employed for learning in the brain, in contrast with original studies which suggested that dopaminergic neurons all report one unitary prediction error signal Schultz et al. (1997). This may not be surprising from a theoretical point of view, as learning about two (or more) separate reward predictions within any given scenario requires the calculation of two separate prediction errors. Such a dual-task situation may be common in daily life. However, reinforcement learning tasks previously examined in laboratory settings did not directly test this prediction.

3.2 Neural correlates of pseudo-reward

A second prediction states that prediction errors should occur in connection to task subgoals as well as with top level goals. HRL agents associate a special form of reward to subgoals, dubbed *pseudo-reward* (Sutton et al., 1999). Distinguishing pseudo-rewards from primary reward is crucial: If subgoals were confounded with primary reward, the agent might get stuck "chasing" subgoals, even when irrelevant to top-level goals. The distinction between pseudo and primary rewards results in a distinction between two types of prediction errors: ordinary reward prediction errors (RPEs) occur in response to differences in predicted outcomes in progress towards primary goals. Pseudo-reward prediction errors (PPEs) occur in response to outcomes in progress towards subgoals. PPEs are unique to HRL, they do not occur in ordinary, "flat" RL. If HRL is relevant

to neural activity, we should expect to see neural correlates of PPEs. Ribas-Fernandes et al. (2011) designed a task to test this prediction, using EEG and fMRI to assay for a neural analogue to the pseudo-reward prediction error.

Figure 2 illustrates the task. 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.

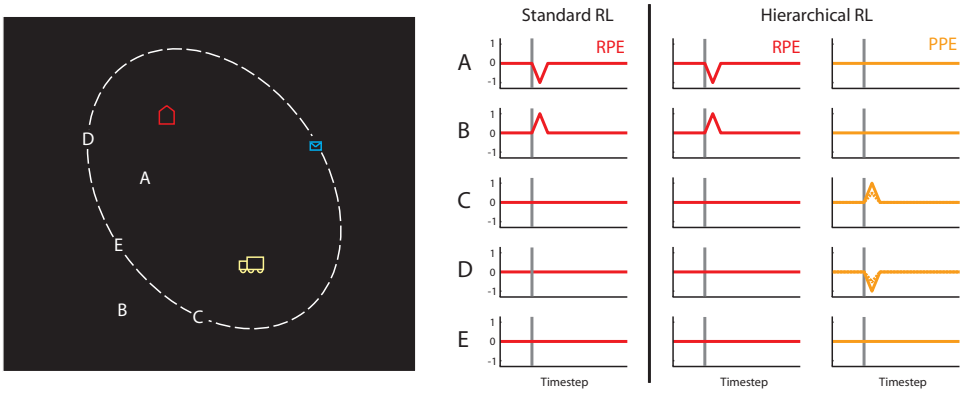


Fig. 2. Task and predictions from HRL and RL (from (Ribas-Fernandes et al., 2011)). Left: Task display and underlying geometry of the delivery task. Right: Prediction-error signals generated by standard RL and by HRL in each category of jump event. Grey bars mark the time-step immediately preceding a jump event. Dashed time-courses indicate the PPE generated in C and D jumps that change the subgoals distance by a smaller amount.

An additional twist was that on some trials, the package unexpectedly jumped to a new location before the truck reached 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.) 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. According to HRL, a jump to this location should thus 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 a RPE nor a PPE). These points are illustrated in the right panel of Figure 2, which shows RPE and PPE time-courses from simulations of the delivery task based on standard RL and HRL.

A group of 30 participants performed the delivery task while undergoing fMRI. Here, one third of the trials included a jump of type D (see Figure 2), predicted to elicit a negative PPE. Neural correlates for such a jump were found in dorsal anterior cingulate cortex and habenula, structures previously suggested to reflect or induce reduced dopaminergic activity.

Because these PPEs are unique to HRL, not occurring in standard RL, Ribas-Fernandes et al. (2011) interpreted these results as providing a neural signature of HRL.

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? This question has added relevance because the wrong set of skills can actually impair learning (Botvinick et al., 2009).

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 motivated to reach an option's subgoal, once the option gains control of behavior. In HRL, as discussed in the previous section, 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, pseudo-reward must be assigned to specific outcomes, 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 portrays options and subgoals 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).

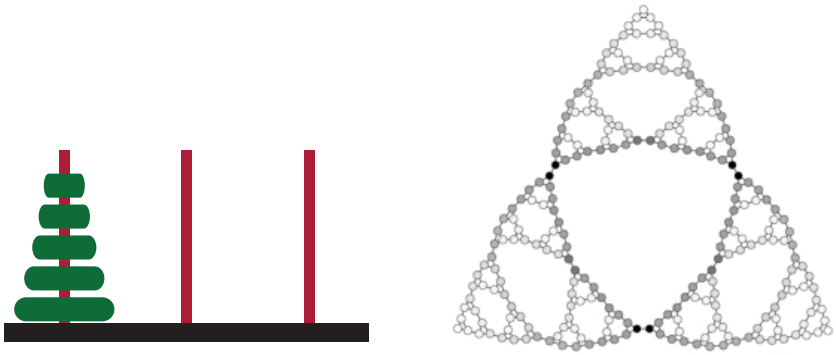
Another approach to explaining subgoal discovery leverages the notion of intrinsic motivation. The idea here is that certain events or stimuli are inherently interesting to the behaving animal or human. These can be stimuli that display salient perceptual properties or that challenge expectations, eliciting curiosity (Schmidhuber, 1991a,b). In an HRL context such states are proposed to be adopted as subgoals, 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. 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 subgoal 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 3.



(a) (b)

Fig. 3. (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 3(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

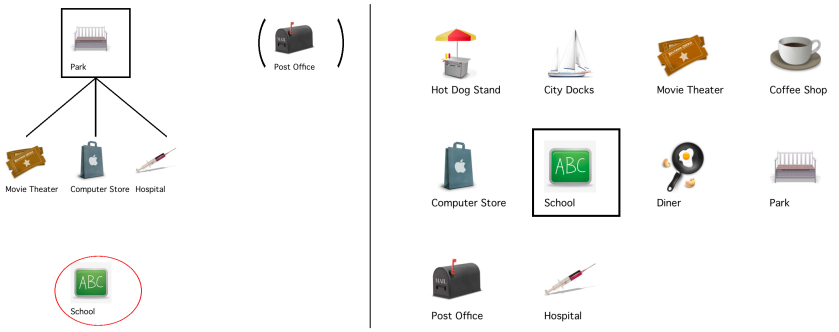
In a first experiment we investigated whether humans can identify bottleneck states, when doing so allows them to optimize their performance. We summarize the experiment and its results here; full details are presented in Diuk et al. (2012a).

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 4(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 5(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.

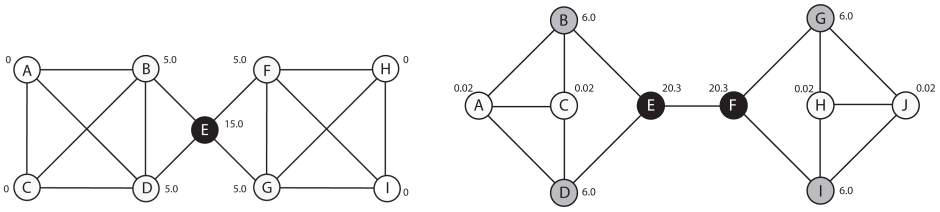
Note that participants were never actually shown the graph in Figure 5(a), or any other sort of birds-eye view of the town. The display only provided information about local adjacencies. Nevertheless, we hypothesized that, with accumulating experience, participants would identify the bottleneck location and exploit it by selecting it as a bus-stop location. Figure 6 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 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.

The results of this experiment do not, however, allow us to make conclusions about *how* participants identified the bottleneck location. In particular, while



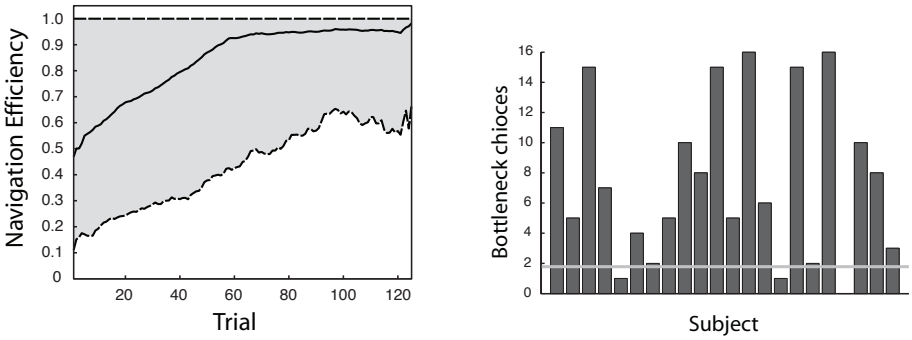
(a) First version. (b) Second version

Fig. 4. 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. 5. 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.



(a) (b)

Fig. 6. (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.

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, reported in detail in Diuk et al. (2012a), removed the confound between centrality and frequency, and used a graph in which all vertices had the same degree (Figure 5(b)). Figure 4(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%.

In a further experiment, which we only briefly summarize here, Diuk et al. (2012a) showed that when participants were given a start and goal location, and asked to verify whether a third location would fall on a shortest delivery path, they were especially fast at responding the question when the probe location corresponded to a domain bottleneck. The finding suggests participants formulated delivery plans using bottleneck locations as subgoals.[1]

Together, the foregoing 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.

[1] This particular result provides preliminary evidence for "model-based" hierarchical planning in the Diuk et al. (2012a) delivery task.

6 Experiment 3: Bottleneck States and Temporal Abstraction

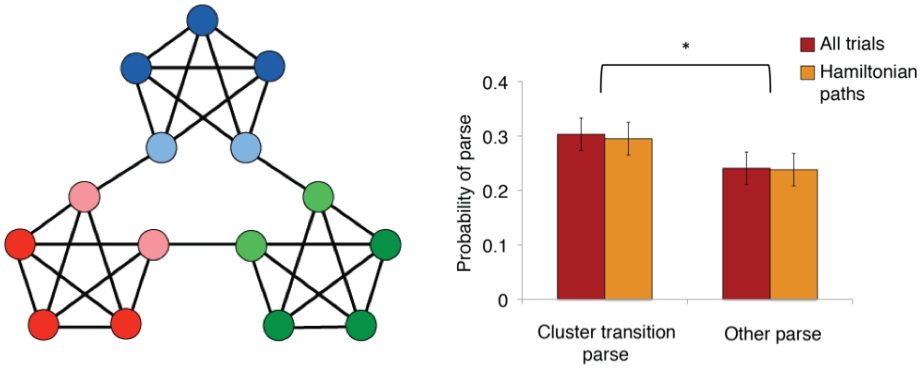
Our approach in Experiment 3 was based on previous work using event parsing. A standard experimental paradigm in cognitive psychology involves showing an action sequence, and asking participants to "parse" it by pressing a key when they feel that one subsequence or subtask has ended and a new one has 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. (2012); we summarize the work here.

Participants were exposed to a sequence of images presented one at a time over a period of 35 minutes. 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 7(a). Each of the 15 possible images 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. Drawing on the complex network literature, we refer to these clusters as "communities" (see Schapiro et al., 2012).

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.

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.

We additionally hypothesized that stimuli grouped together as part of the same event on the basis of community structure might come to be represented more similarly in the brain. Participants in a new experiment were exposed to sequences of stimuli generated from the graph in Figure 7 in the fMRI scanner. To test the hypothesis that items in the same community would be represented more similarly, we analyzed the similarity of the patterns of voxel activation evoked by the stimuli in searchlights throughout the brain. We found that the patterns of activation for items within a community were more similar than



(a) (b)

Fig. 7. (a) Underlying graph of the task. Each node in the graph is linked to a stimulus used in the sequence. (b) Proportion of times participants parsed sequence at cluster-changing points, as opposed to other points in the sequence.

those for items between communities in left IFG, anterior temporal lobe (ATL), insula, and superior temporal gyrus (STG) (see Figure 8).

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 consistent 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

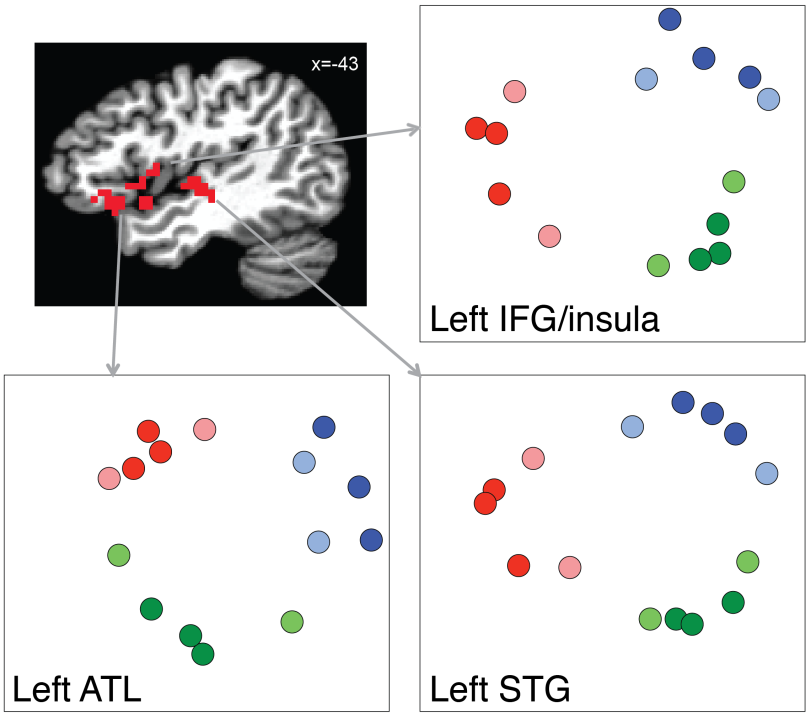


Fig. 8. Pattern similarity effects in left IFG/insula, left ATL, and left STG. Each cluster showed reliable community structure in the BOLD response in a whole-brain search. The similarity structure within each area is visualized using multi-dimensional scaling, with items color-coded in accordance with the graph nodes in Figure 7

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.

One outstanding question is how subtasks are transferred, once learned. Even though many hierarchical problems share exactly matching subtasks (boiling water for the preparation of both tea and coffee), many other problems faced by humans and animals have only partially overlapping states or actions. A richer understanding of subtask learning should include a mechanism for such less constrained transfer.

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.

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.

Dayan, P. and Hinton, G. E. (1993). Feudal reinforcement learning. In *Advances in Neural Information Processing Systems 5*, pages 271–278. Morgan Kaufmann.

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

Diuk, C., Cordova, N., Niv, Y., and Botvinick, M. (2012a). Discovering hierarchical task structure. *Submitted*.

Diuk, C., Tsai, K., Wallis, J., Niv, Y., and Botvinick, M. M. (2012b). Two simultaneous, but separable, reward prediction errors in human ventral striatum. *Submitted*.

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.

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 networks : the official journal of the International Neural Network Society*, 19(8):1242–54.

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

Ito, M. and Doya, K. (2011). Multiple representations and algorithms for reinforcement learning in the cortico-basal ganglia circuit. *Current opinion in neurobiology*, 21(3):368–73.

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., Solway, A., Diuk, C., McGuire, J. T., Barto, A. G., Niv, Y., and Botvinick, M. M. (2011). A neural signature of hierarchical reinforcement learning. *Neuron*, 71(2):370–9.

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. (2012). Neural representations of events arise from temporal community structure. *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.