

Hierarchically organized behavior and its neural foundations: A reinforcement learning perspective

Matthew M. Botvinick^{a,*}, Yael Niv^a, Andrew G. Barto^b

^a Princeton Neuroscience Institute, Department of Psychology, Princeton University, Green Hall, Princeton, NJ 08540, United States

^b University of Massachusetts, Amherst, Department of Computer Science, United States

ARTICLE INFO

Article history:

Received 11 October 2007

Revised 12 August 2008

Accepted 24 August 2008

Keywords:

Reinforcement learning
Prefrontal cortex

ABSTRACT

Research on human and animal behavior has long emphasized its hierarchical structure—the divisibility of ongoing behavior into discrete tasks, which are comprised of subtask sequences, which in turn are built of simple actions. The hierarchical structure of behavior has also been of enduring interest within neuroscience, where it has been widely considered to reflect prefrontal cortical functions. In this paper, we reexamine behavioral hierarchy and its neural substrates from the point of view of recent developments in computational reinforcement learning. Specifically, we consider a set of approaches known collectively as *hierarchical reinforcement learning*, which extend the reinforcement learning paradigm by allowing the learning agent to aggregate actions into reusable subroutines or skills. A close look at the components of hierarchical reinforcement learning suggests how they might map onto neural structures, in particular regions within the dorsolateral and orbital prefrontal cortex. It also suggests specific ways in which hierarchical reinforcement learning might provide a complement to existing psychological models of hierarchically structured behavior. A particularly important question that hierarchical reinforcement learning brings to the fore is that of how learning identifies new action routines that are likely to provide useful building blocks in solving a wide range of future problems. Here and at many other points, hierarchical reinforcement learning offers an appealing framework for investigating the computational and neural underpinnings of hierarchically structured behavior.

© 2008 Elsevier B.V. All rights reserved.

1. Introduction

In recent years, it has become increasingly common within both psychology and neuroscience to explore the applicability of ideas from machine learning. Indeed, one can now cite numerous instances where this strategy has been fruitful. Arguably, however, no area of machine learning has had as profound and sustained an impact on psychology and neuroscience as that of computational reinforcement learning (RL). The impact of RL was initially felt in research on classical and instrumental conditioning (Barto & Sutton, 1981; Sutton & Barto, 1990; Wickens,

Kotter, & Houk, 1995). Soon thereafter, its impact extended to research on midbrain dopaminergic function, where the temporal-difference learning paradigm provided a framework for interpreting temporal profiles of dopaminergic activity (Barto, 1995; Houk, Adams, & Barto, 1995; Montague, Dayan, & Sejnowski, 1996; Schultz, Dayan, & Montague, 1997). Subsequently, actor–critic architectures for RL have inspired new interpretations of functional divisions of labor within the basal ganglia and cerebral cortex (see Joel, Niv, and Ruppel (2002) for a review), and RL-based accounts have been advanced to address issues as diverse as motor control (e.g., Miyamoto, Morimoto, Doya, & Kawato, 2004), working memory (e.g., O'Reilly & Frank, 2006), performance monitoring (e.g., Holroyd & Coles, 2002), and the distinction between

* Corresponding author. Tel.: +1 609 258 1280.

E-mail address: matthewwb@princeton.edu (M.M. Botvinick).

habitual and goal-directed behavior (e.g., Daw, Niv, & Dayan, 2005).

As ideas from RL permeate the fields of psychology and neuroscience, it is interesting to consider how RL research has continued to evolve within computer science. Here, attention has turned increasingly to factors that limit the applicability of RL. Perhaps foremost among these is the *scaling problem*: Unfortunately, basic RL methods do not cope well with large task domains, i.e., domains involving a large space of possible world states or a large set of possible actions. This limitation of RL has been little discussed within psychology and neuroscience, where RL has typically been applied to highly simplified learning situations. However, the scaling problem has direct implications for whether RL mechanisms can be plausibly applied to more complex behavioral contexts. Because such contexts would naturally include most scenarios animals and human beings face outside the laboratory, the scaling problem is clearly of relevance to students of behavior and brain function.

A number of computational approaches have been developed to tackle the scaling problem. One increasingly influential approach involves the use of *temporal abstraction* (Barto & Mahadevan, 2003; Dietterich, 2000; Parr & Russell, 1998; Sutton, Precup, & Singh, 1999). Here, the basic RL framework is expanded to include temporally abstract actions, representations that group together a set of interrelated actions (for example, grasping a spoon, using it to scoop up some sugar, moving the spoon into position over a cup, and depositing the sugar), casting them as a single higher-level action or skill ('add sugar'). These new representations are described as temporal abstractions because they abstract over temporally extended, and potentially variable, sequences of lower-level steps. A number of other terms have been used as well, including 'skills,' 'operators,' 'macro-operators' and 'macro-actions.' In what follows, we will often refer to temporally abstract actions as *options*, following Sutton et al. (1999).

In most versions of RL that use temporal abstraction, it is assumed that options can be assembled into higher-level skills in a hierarchical arrangement. Thus, for example, an option for adding sugar might form part of other options for making coffee and tea. Given the importance of such hierarchical structures in work using temporal abstraction, this area of RL is customarily referred to as hierarchical reinforcement learning (HRL).

The emergence of HRL is an intriguing development from the points of view of psychology and neuroscience, where the idea of hierarchical structure in behavior is familiar. In psychology, hierarchy has played a pivotal role in research on organized, goal-directed behavior, from the pioneering work in this area (e.g., Estes, 1972; Lashley, 1951; Miller, Galanter, & Pribram, 1960; Newell & Simon, 1963) through to the most recent studies (e.g., Anderson, 2004; Botvinick & Plaut, 2004; Schneider & Logan, 2006; Zacks, Speer, Swallow, Braver, & Reynolds, 2007). Behavioral hierarchy has also been of longstanding interest within neuroscience, where it has been considered to relate closely to prefrontal cortical function (Badre, 2008; Botvinick, 2008; Courtney, Roth, & Sala, 2007; Fuster,

1997; Koehlin, Ody, & Kouneiher, 2003; Wood & Grafman, 2003).

Thus, although HRL was not originally developed to address questions about human and animal behavior, it is potentially of twofold relevance to psychology and neuroscience. First, HRL addresses a limitation of RL that would also be faced by any biological agent learning through RL-like mechanisms. The question thus naturally arises whether the brain might deal with this limitation in an analogous way. Second, the ideas at the heart of HRL resonate strongly with existing themes in psychology and neuroscience. The formal framework provided by HRL thus might provide leverage in thinking about the role of hierarchical structure in human and animal behavior, and in particular how such structure might relate to behavioral and neuroscientific issues that have already been treated in terms of RL.

Our objective in the present paper is to consider HRL from these two perspectives. We begin, in the following section, by examining the scaling problem and considering how the use of temporal abstraction can help to ameliorate it. We then turn to HRL itself, detailing its representational and algorithmic assumptions. After establishing these, we discuss the potential implications of HRL for behavioral research. Here, we emphasize one fundamental computational issue that HRL brings into focus, which concerns the question of how reusable sets of skills might develop through learning. Finally, we consider the potential implications of HRL for interpreting neural function. To this end, we introduce a new actor–critic implementation of HRL, which makes explicit the computational requirements that HRL would pose for a neural implementation.

2. Temporal abstraction and the scaling problem

A key source of the scaling problem is the fact that an RL agent can learn to behave adaptively only by exploring its environment, trying out different courses of action in different situations or states of the environment, and sampling their consequences. As a result of this requirement, the time needed to arrive at a stable behavioral policy increases with both the number of different states in the environment and the number of available actions. In most contexts, the relationship between training time and the number of environmental states or actions is a positively accelerating function. Thus, as problem size increases, standard RL eventually becomes infeasible.

Numerous approaches have been adopted in machine learning to deal with the scaling problem. These include reducing the size of the state space by suppressing behaviorally irrelevant distinctions between states (state abstraction; see, e.g., Li and Walsh (2006)), and methods aimed at striking an optimal balance between exploration and exploitation of established knowledge (e.g., Kearns & Singh, 2002). HRL methods target the scaling problem by introducing temporally abstract actions (Barto & Mahadevan, 2003; Dietterich, 2000; Parr & Russell, 1998; Sutton et al., 1999). The defining characteristic of these abstract actions is that, rather than specifying a single 'primitive' action to execute, each abstract action instead

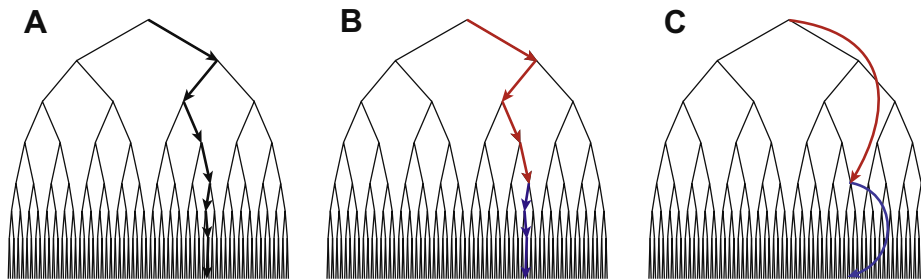


Fig. 1. An illustration of how options can facilitate search. (A) A search tree with arrows indicating the pathway to a goal state. A specific sequence of seven independently selected actions is required to reach the goal. (B) The same tree and trajectory, the colors indicating that the first four and the last three actions have been aggregated into options. Here, the goal state is reached after only two independent choices (selection of the options). (C) Illustration of search using option models, which allow the ultimate consequences of an option to be forecast without requiring consideration of the lower-level steps that would be involved in executing the option. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

specifies a whole *policy* to be followed, that is, a mapping from states to actions.¹ Once a temporally abstract action is initiated, execution of its policy continues until any member of a set of specified termination states is reached.² Thus, the selection of a temporally abstract action ultimately results in the execution of a sequence of primitive actions.

Adding temporal abstraction to RL can ease the scaling problem in two ways. The first way is through its impact on the exploration process. In order to see how this works, it is useful to picture the agent (i.e., the simulated human or animal) as searching a tree structure (Fig. 1A). At the apex is a node representing the state occupied by the agent at the outset of exploration. Branching out from this node are links representing primitive actions, each leading to a node representing the state (and, possibly, reward) consequent on that action. Further action links project from each of these nodes, leading to their consequent states, and so forth. The agent's objective is to discover paths through the decision tree that lead to maximal accumulated rewards. However, the set of possible paths increases with the set of actions available to the agent, and with the number of reachable states. With increasing numbers of either it becomes progressively more difficult to discover, through exploration, the specific traversals of the tree that would maximize reward.

Temporally abstract actions can alleviate this problem by introducing *structure* into the exploration process. Specifically, the policies associated with temporally abstract actions can guide exploration down specific partial paths through the search tree, potentially allowing earlier discovery of high-value traversals. The principle is illustrated in Fig. 1A and B. Discovering the pathway illustrated in Fig. 1A using only primitive, one-step actions, would require a specific sequence of seven independent choices. This changes if the agent has acquired—say, through prior experience with related problems—two options corresponding to the differently colored subsequences in Fig. 1B. Equipped with these, the agent would only need

to make two independent decisions to discover the overall trajectory, namely, select the two options. Here, options reduce the effective size of the search space, making it easier for the agent to discover an optimal trajectory.

The second, and closely related, way in which temporally abstract actions can ease the scaling problem is by allowing the agent to learn more efficiently from its experiences. Without temporal abstraction, learning to follow the trajectory illustrated in Fig. 1A would involve adjusting parameters at seven separate decision-points. With predefined options (Fig. 1B), policy learning is only required at two decision-points, the points at which the two options are to be selected. Thus, temporally abstract actions not only allow the agent to explore more efficiently, but also to make better use of its experiences.

Along with these advantages, there also comes a new computational burden. For in order to enjoy the benefits of temporal abstraction, the agent must have some way of *acquiring* a set of *useful* options. As we shall discuss, this requirement raises some of the most interesting issues in HRL, issues that also apply to human learning.

3. Hierarchical reinforcement learning

Having briefly discussed the motivation for incorporating temporal abstraction into RL, we now turn to a more direct description of how HRL operates. For simplicity, we focus on one specific implementation of HRL, the *options framework* described by Sutton et al. (1999). However, the points we shall emphasize are consistent with other versions of HRL, as well (e.g., Dietterich (2000) and Parr and Russell (1998); for an overview, see Barto and Mahadevan (2003)). Since one of our objectives is to explore potential neuroscientific correlates of HRL, we have implemented the options framework within an actor–critic architecture (defined below), allowing direct parallels to be drawn with previous work relating RL to functional neuroanatomy through the actor–critic framework.³ In what follows, we provide an informal, tutorial-style overview of

¹ An alternative term for temporal abstraction is thus *policy abstraction*.

² Some versions of HRL allow for options to be interrupted at points where another option or action is associated with a higher expected value. See, e.g., Sutton et al. (1999).

³ For other work translating HRL into an actor–critic format, see Bhatnagara and Panigrahi (2006).

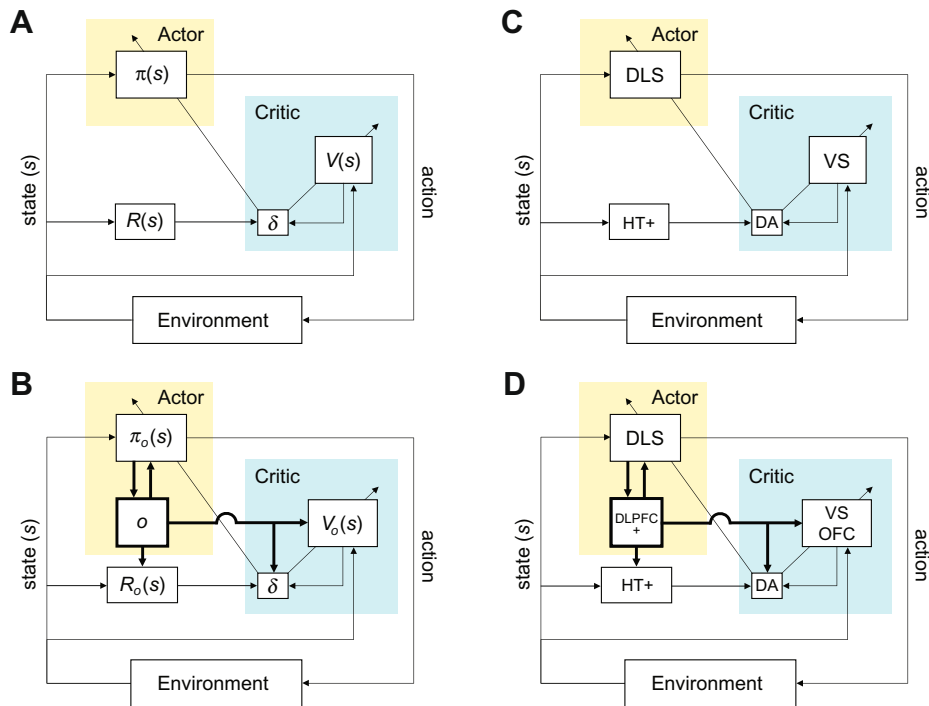


Fig. 2. An actor–critic implementation. (A) Schematic of the basic actor–critic architecture. $R(s)$: reward function; $V(s)$: value function; δ : temporal-difference prediction error; $\pi(s)$: policy, determined by action strengths W . (B) An actor–critic implementation of HRL. o : currently controlling option, $R_o(s)$: option-dependent reward function. $V_o(s)$: option-specific value functions; δ : temporal-difference prediction error; $\pi_o(s)$: option-specific policies, determined by option-specific action/option strengths. (C) Putative neural correlates to components of the elements diagrammed in panel A. (D) Potential neural correlates to components of the elements diagrammed in panel A. (D) Potential neural correlates to components of the elements diagrammed in panel A. Abbreviations: DA: dopamine; DLPFC: dorsolateral prefrontal cortex, plus other frontal structures potentially including premotor, supplementary motor and pre-supplementary motor cortices; DLS: dorsolateral striatum; HT+: hypothalamus and other structures, potentially including the habenula, the pedunculopontine nucleus, and the superior colliculus; VS: ventral striatum.

this implementation. Full technical details are presented in the [Appendix](#).

3.1. Fundamentals of RL: temporal-difference learning in actor–critic models

RL problems comprise four elements: a set of world *states*; a set of *actions* available to the agent in each state; a *transition function*, which specifies the probability of transitioning from one state to another when performing each action; and a *reward function*, which indicates the amount of reward (or cost) associated with each such transition. Given these elements, the objective for learning is to discover a policy, that is, a mapping from states to actions, that maximizes cumulative long-term reward.⁴

In actor–critic implementations of RL, the learning agent is divided into two parts, an actor and a critic, as illustrated in Fig. 2A (see, e.g., Barto, Sutton, & Anderson, 1983; Houk et al., 1995; Joel et al., 2002; Suri, Bargas, & Arbib, 2001). The *actor* selects actions according to a modifiable policy ($\pi(s)$ in the figure), which is based on a set of

weighted associations from states to actions, often called *action strengths*. The *critic* maintains a *value function* ($V(s)$), associating each state with an estimate of the cumulative, long-term reward that can be expected subsequent to visiting that state. Importantly, both the action strengths and the value function must be learned based on experience with the environment. At the outset of learning, the value function and the actor’s action strengths are initialized, for instance uniformly or randomly, and the agent is placed in some initial state. The actor then selects an action, following a rule that favors high-strength actions but also allows for exploration (see [Appendix](#), Eq. (1)). Once the resulting state is reached and its associated reward is collected, the critic computes a *temporal-difference prediction error* (denoted δ in the figure; see also Eq. (2)). Here, the value that was attached to the previous state is treated as a prediction of (1) the reward that would be received in the successor state ($R(s)$), plus (2) the value attached to that successor state. A positive prediction error indicates that this prediction was too low, meaning that things turned out better than expected. Of course, things can also turn out worse than expected, yielding a negative prediction error.

The prediction error is used to update both the value attached to the previous state and the strength of the action that was selected in that state (see Eqs. (3) and (4)). A positive prediction error leads to an increase in the value of

⁴ It is often assumed that the utility attached to rewards decreases with the length of time it takes to obtain them, and in such cases the objective is to maximize the *discounted* long-term reward. As reflected in the [Appendix](#), our implementation assumes such discounting. For simplicity, however, discounting is ignored in the main text.

the previous state and the propensity to perform the chosen action at that state. A negative error leads to a reduction in these. After the appropriate adjustments, the agent selects a new action, a new state is reached, a new prediction error is computed, and so forth. As the agent explores its environment and this procedure is repeated, the critic's value function becomes progressively more accurate, and the actor's action strengths change so as to yield progressive improvements in behavior, in terms of the amount of reward obtained.

3.2. Incorporating temporally abstract actions

The options framework supplements the set of single-step, primitive actions with a set of temporally abstract actions or options. An option is, in a sense, a 'mini-policy.' It is defined by an *initiation set*, indicating the states in which the option can be selected; a *termination function*, which specifies a set of states that will trigger termination of the option⁵; and an *option-specific policy*, mapping from states to actions (which now include other options).

Like primitive actions, options are associated with strengths, and on any time-step the actor may select either a primitive action or an option. Once an option is selected, actions are selected based on that option's policy until the option terminates. At that point, a prediction error for the option is computed. This error is defined as the difference between the value of the state where the option terminated and the value of the state where the option was initiated, plus whatever rewards were accrued during execution of the option (see Eq. (6)). A positive prediction error indicates that things went better than expected since leaving the initiation state, and a negative prediction error means that things went worse. As in the case of primitive actions, the prediction error is used to update the value associated with the initiation state, as well as the action strength associating the option with that state (see Eqs. (8) and (9); Fig. 3).⁶

Implementing this new functionality requires several extensions to the actor-critic architecture, as illustrated in Fig. 2B. First, the actor must maintain a representation of which option is currently in control of behavior (o).⁷ Second, because the agent's policy now varies depending on which option is in control, the actor must maintain a separate set of action strengths for each option ($\pi_o(s)$ in the figure). Important changes are also required in the critic. Because prediction errors are computed when options terminate, the critic must receive input from the actor, telling it when such terminations occur (the arrow from o to δ). Finally, to be able to compute the prediction error at these points, the critic must also keep track of the amount of re-

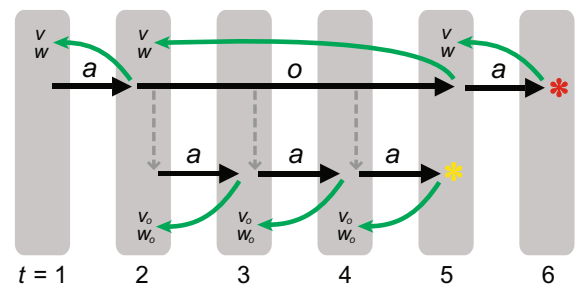


Fig. 3. A schematic illustration of HRL dynamics. a , primitive actions; o , option. On the first time-step ($t = 1$), the agent executes a primitive action (forward arrow). Based on the consequent state (i.e., the state at $t = 2$), a prediction error δ is computed (arrow running from $t = 2$ to $t = 1$), and used to update the value (V) and action/option strengths (W) associated with the preceding state. At $t = 2$, the agent selects an option (long forward arrow), which remains active through $t = 5$. During this time, primitive actions are selected according to the option's policy (lower tier of forward arrows). A prediction error is computed after each (lower tier of curved arrows), and used to update the option-specific values (V_o) and action strengths (W_o) associated with the preceding state. These prediction errors, unlike those at the level above, take into account pseudo-reward received throughout the execution of the option (higher asterisk). Once the option's subgoal state is reached, the option is terminated. A prediction error is computed for the entire option (long curved arrow), and this is used to update the values and option strengths associated with the state in which the option was initiated. The agent then selects a new action at the top-level, which yields external reward (lower asterisk). The prediction errors computed at the top-level, but not at the level below, take this reward into account.

ward accumulated during each option's execution and the identity of the state in which the option was initiated (see Eqs. (6)–(9)).

3.3. Learning option policies

The description provided so far explains how the agent learns a top- or root-level policy, which determines what action or option to select when no option is currently in control of behavior. We turn now to the question of how option-specific policies are learned.

In versions of the options framework that address such learning, it is often assumed that options are initially defined in terms of specific *subgoal* states. (The question of where these subgoals come from is an important one, which we address later.) It is further assumed that when an active option reaches its subgoal, the actions leading up to the subgoal are reinforced. To distinguish this reinforcing effect from the one associated with external rewards, subgoal attainment is said to yield *pseudo-reward* (Dietterich, 2000).

In order for subgoals and pseudo-reward to shape option policies, the critic in HRL must maintain not only its usual value function, but also a set of option-specific value functions ($V_o(s)$ in Fig. 2B). As in ordinary RL, these value functions predict the cumulative long-term reward that will be received subsequent to occupation of a particular state. However, they are option-specific in the sense that they take into account the pseudo-reward that is associated with each option's subgoal state. A second reason that option-specific value functions are needed is that the reward (and pseudo-reward) that the agent will receive

⁵ The termination function may be probabilistic.

⁶ As discussed by Sutton et al. (1999), it is possible to update the value function based only on comparisons between states and their immediate successors. However, the relevant procedures, when combined with those involved in learning option-specific policies (as described later), require complicated bookkeeping and control operations for which neural correlates seem less plausible.

⁷ If it is assumed that option policies can call other options, then the actor must also keep track of the entire set of active options and their calling relations.

following any given state depends on the actions it will select. These depend, by definition, on the agent's policy, and under HRL the policy depends on which option is currently in control of behavior. Thus, only an option-specific value function can accurately predict future rewards.

Despite the additions above, option-specific policies are learned in quite the usual way: On each step of an option's execution, a prediction error is computed based on the (option-specific) values of the states visited and the reward received (including pseudo-reward). This prediction error is then used to update the option's action strengths and the values attached to each state visited during the option (see Eqs. (6)–(9); Fig. 3). With repeated cycles

through this procedure, the option's policy evolves so as to guide behavior, with increasing directness, toward the option's subgoals.

3.4. Illustrations of performance

To provide an illustration of HRL in action, we applied the preceding learning procedures to a toy 'rooms' problem introduced by Sutton et al. (1999). Here, the agent's task is to navigate through a set of rooms interconnected by doorways, in order to reach a goal state (Fig. 4A). In each state, the agent can select any of eight deterministic primitive actions, each of which moves the agent to one of the adja-

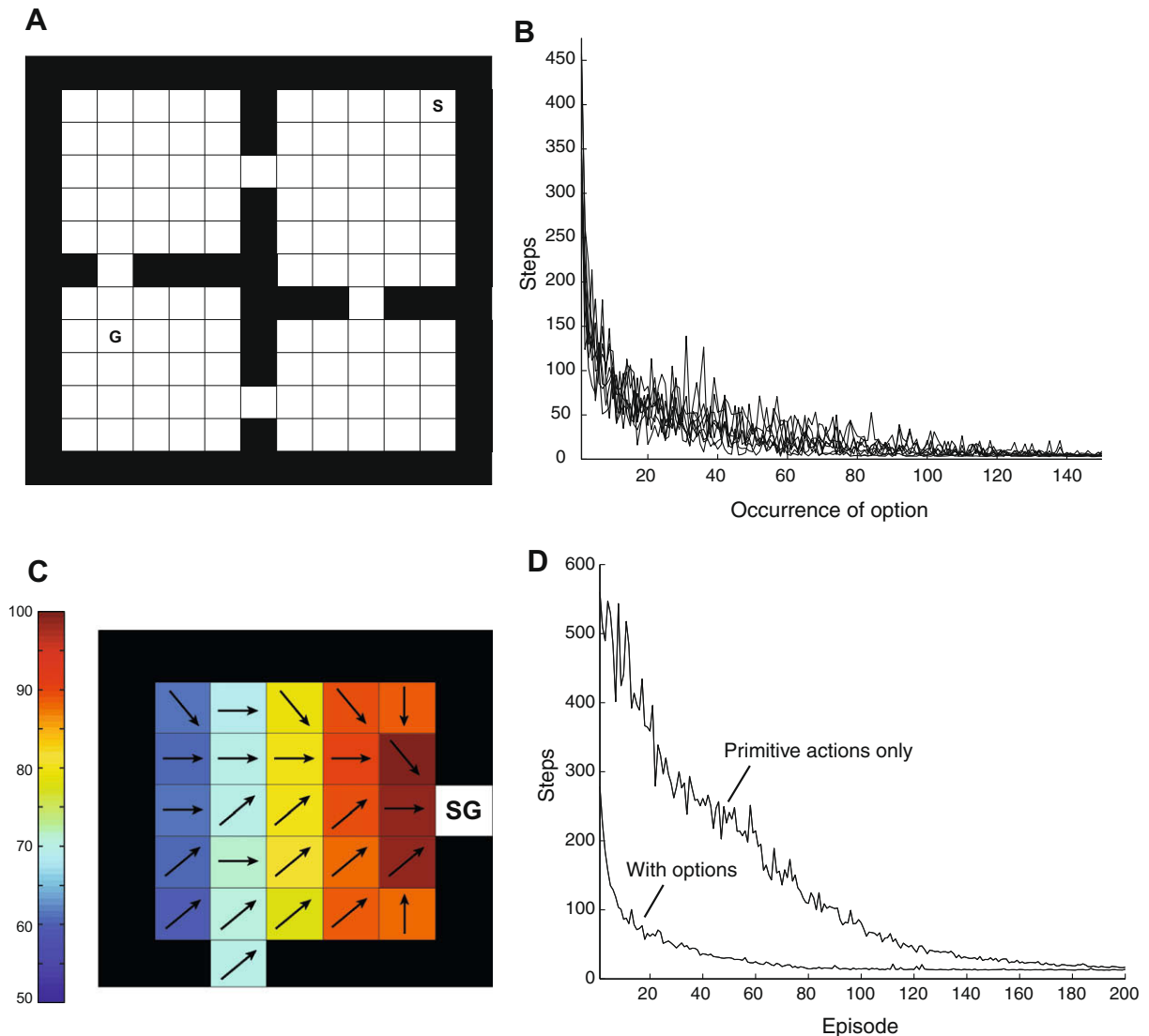


Fig. 4. (A) The rooms problem, adapted from Sutton et al. (1999). S: start; G: goal. (B) Learning curves for the eight doorway options, plotted over the first 150 occurrences of each (mean over 100 simulation runs). See Appendix for simulation details. (C) The upper left room from panel A, illustrating the policy learned by one doorway option. Arrows indicate the primitive action selected most frequently in each state. SG: option subgoal. Colors indicate the option-specific value for each state. (D) Learning curves indicating solution times, i.e., number of primitive steps to goal, on the problem illustrated in panel A (mean over 100 simulation runs). Upper data series: performance when only primitive actions were included. Lower series: performance when both primitive actions and doorway options were included. Policies for doorway options were established through earlier training (see Appendix). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

cent squares (unless a wall prevents this movement). Additionally, within each room the agent can also select either of two options, each having one of the room's doors as its subgoal.

To illustrate the process of learning option-specific policies, the model was initially trained with only pseudo-rewards at the option subgoal states, i.e., without external reward. Fig. 4B tracks the number of primitive actions each option required to reach its subgoal, showing that, through learning, this fell to a minimum over successive executions of the option. Fig. 4C illustrates the policy learned by one of the doorway options, as well its option-specific value function.

A more fundamental point is illustrated in Fig. 4D, which tracks the model's performance after external rewards were introduced at the goal state *G*. The model learns more rapidly to reach the goal state when both the doorway options and the eight primitive actions are included⁸ than when only the primitive actions are available. This savings in training time reflects the impact of temporal abstraction on exploration and learning, as described in the previous section.

3.5. Other versions of HRL

As noted earlier, we based the foregoing overview of HRL on a particular implementation of HRL, the options framework (Sutton et al., 1999). Although we continue to orient toward this paradigm throughout the rest of the article, it is important to bear in mind that there exist other versions of HRL, which differ from the options framework along a number of dimensions (see Barto & Mahadevan, 2003). The most highly developed among these are the HAM framework introduced by Parr and Russell (1998) and Dietterich's (1998, 2000) MAXQ framework. One aspect that distinguishes these two paradigms from the options framework is that they treat learning as a process occurring within a pre-established and partially fixed task/subtask hierarchy.⁹ This approach has ramifications that further differentiate the HAM and MAXQ paradigms from the options framework, affecting, for example, the way that value functions are represented and how prediction errors are computed (see Barto and Mahadevan (2003) for a detailed comparison among approaches).

We have chosen to focus on the options framework because the elements that it adds to standard RL are simpler and fewer in number than those added by other paradigms, and because in our opinion these new elements lend themselves more immediately to a neuroscientific interpretation. In the General Discussion, we consider several ways in which other HRL paradigms might lead to subtly differ-

ent hypotheses and predictions. Nevertheless, we endeavor in what follows to concentrate on points that are broadly consistent with the full range of HRL implementations.

4. Behavioral implications

Having introduced the fundamentals of HRL, we turn now to a consideration of what their implications might be for behavioral and neuroscientific research. We begin with implications for psychology. As noted earlier, HRL treats a set of issues that have also been of longstanding interest to students of human and animal behavior. HRL suggests a different way of framing some of these issues, and also brings to the fore some important questions that have so far received relatively little attention in behavioral research.

4.1. Relation to previous work in psychology

Lashley (1951) is typically credited with first asserting that the sequencing of low-level actions requires higher-level representations of task context. Since this point was introduced, there has been extensive research into the nature and dynamics of such representations, much of which resonates with the idea of temporally abstract actions as found in HRL. Indeed, the concept of 'task representation,' as it arises in much contemporary psychological work (e.g., Cohen, Dunbar, & McClelland, 1990; Cooper & Shallice, 2000; Monsell, 2003), shares key features with the option construct. Both postulate a unitary representation that (1) can be selected or activated; (2) remains active for some period of time following its initial selection; (3) leads to the imposition of a specific stimulus-response mapping or policy; and (4) can participate in hierarchical relations with other representations of the same kind.

Despite this parallel, most psychological research on task representation has focused on issues different from those central to HRL. In recent work, the emphasis has often been on the dynamics of shifts from one task to another (e.g., Allport & Wylie, 2000; Logan, 2003; Monsell, 2003), or on competition between task sets (e.g., Monsell, Yeung, & Azuma, 2000; Pashler, 1994). Other studies have concentrated on cases where task representations function primarily to preserve information conveyed by transient cues (e.g., Cohen, Braver, & O'Reilly, 1996; MacDonald, Cohen, Stenger, & Carter, 2000), a function not usually performed by options.

Among studies focusing on the issue of hierarchy, many have aimed at obtaining empirical evidence that human behavior and its accompanying mental representations are in fact organized in a hierarchical fashion (e.g., Newton, 1976; Zacks & Tversky, 2001). However, there have also been a series of theoretical proposals concerning the control structures underlying hierarchically organized behavior (e.g., Arbib, 1985; Botvinick & Plaut, 2004; Cooper & Shallice, 2000; Dehaene & Changeux, 1997; Dell, Berger, & Svec, 1997; Estes, 1972; Grossberg, 1986; MacKay, 1987; Miller et al., 1960; Rumelhart & Norman, 1982). The resemblance between these proposals and HRL mecha-

⁸ As detailed in the Appendix, options were pre-trained, simulating transfer of knowledge from earlier experience. In this particular problem domain—although not necessarily in general—including options without pretraining slows initial learning on the top-level problem, but later confers a benefit, allowing the agent to converge on an optimal policy earlier than it does in the absence of options.

⁹ Procedures for inducing the initial hierarchy from experience have been explored in MAXQ (see, e.g., Mehta, Ray, Tadepalli, & Dietterich, 2008), but this form of learning is treated as separate from the learning of policies over the hierarchy.

nisms is variable. In most cases, for example, high-level task representations have been understood to send top-down activation directly to action representations, rather than to favor specific links from *stimuli* to responses, as in HRL (however, see Botvinick & Plaut, 2004; Ruh, 2007). Furthermore, in the vast majority of cases the focus has been on aspects of steady-state performance, such as reaction times and error patterns, rather than on the role of temporal abstraction in learning, the focus in HRL.

Having made this latter generalization, it is also important to note several cases in which the role of task representations and hierarchical structure during learning have been directly considered. On the empirical side, there have been a number of studies examining the development of hierarchical structure in the behavior of children (e.g., Bruner, 1973; Fischer, 1980; Greenfield, Nelson, & Saltzman, 1972; Greenfield & Schneider, 1977). The general conclusion of such studies is that, over the course of childhood, behavior shows a hierarchical development, according to which simple operations are gradually incorporated into larger wholes. The fit between this observation and the basic premises of HRL is, of course, clear.

The strongest parallels to HRL within psychology, however, are found in production-system based theories of cognition, in particular Soar (Lehman, Laird, & Rosenbloom, 1996) and ACT-R (Anderson, 2004). A key idea in both of these frameworks is that planning or problem solving can leverage *chunks*, ‘if-then’ rules that can trigger the execution of extended action sequences (Laird, Rosenbloom, & Newell, 1986; Lee & Taatgen, 2003; see also Hayes-Roth & Hayes-Roth, 1979; Ward & Allport, 1997). Like temporally abstract actions in HRL, chunks can facilitate problem solving, increasing the speed and efficiency with which solutions are found. This function allows chunking to provide a natural account for the behavioral phenomenon of *positive transfer*, where improvements in problem-solving efficiency are observed on target problems when these are presented after prior exposure to structurally similar problems.

One factor that differentiates HRL from the Soar and ACT-R frameworks is its organization around the single objective of reward maximization. This aspect of HRL allows it to specify precisely what it means for hierarchically structured behavior to be optimal, and this optimality criterion gives coherence to the learning and performance algorithms involved in HRL (even in cases—encountered regularly, in practice—where HRL does not yield perfect reward-maximizing performance). In contrast, neither ACT-R nor Soar takes reward maximization as a central organizing principle. ACT-R does include ‘production utilities,’ which represent the probability that a given production will lead to achievement of the currently held goal (Anderson, 2004), a feature that resonates with the impact of pseudo-reward in HRL. And there have been recent efforts to integrate RL methods into the Soar framework (Nason & Laird, 2005). Notwithstanding these caveats, the centrality of reward maximization in HRL remains distinctive. A countervailing strength of Soar, ACT-R and related models is that they address a wide range of psychological issues—in particular, limitations in processing capacity—that are not addressed in existing formulations of HRL. The strengths of the two approaches thus appear to be

complementary, and it is exciting to consider ways in which they might be integrated (see Nason and Laird (2005), for some preliminary discussion along these lines).

4.2. Negative transfer

The previous section touched on the phenomenon of positive transfer, where established procedural knowledge facilitates the discovery of solutions to new problems. This phenomenon provides a direct point of contact between human behavior and HRL, where, as demonstrated earlier, options arising from earlier experience can have the same facilitatory effect. However, the literature on transfer effects also highlights a contrary point that pertains equally to HRL, which is that in some circumstances pre-existing knowledge can hinder problem solving. Such *negative transfer* was most famously demonstrated by Luchins (1942), who found that human subjects were less successful at solving word problems when the subjects were first exposed to problems demanding a different solution strategy (see also Landrum, 2005; Rayman, 1982).

A direct analog to negative transfer occurs in HRL when the temporally abstract actions available to the agent are not well suited to the learning problem. For illustration, consider the four-rooms problem described above (see Fig. 4A). However, instead of the doorway options included in the earlier simulation, assume that the agent has a set of options whose subgoals are the states adjacent to the ‘windows’ marked in Fig. 5A. Those options, which are not helpful in solving the problem of reaching the goal state G, cause the agent to spend time exploring suboptimal trajectories, with the effect that learning is slowed overall (Fig. 5B). A subtler but equally informative case is illustrated in Fig. 5C. Here, the original doorway options are used, but now a new passageway has been opened up, providing a shortcut between the upper right and lower left rooms. When trained with primitive actions only, the agent learns to use this passage, finding the shortest path to the reward on 75% of training runs. However, when the original doorway options are also included, the agent learns to reach the goal only by way of the main doorways, eventually ignoring the passageway completely.¹⁰

These illustrations show that the impact of temporally abstract actions on learning and planning depends critically on which specific actions the agent has in its repertoire. This raises a pivotal question, which motivates a significant portion of current HRL research: By what means can a learning agent acquire temporally abstract actions that are likely to be useful in solving future problems, and avoid acquiring unhelpful ones? The existence of both

¹⁰ Mean solution times over the last 10 episodes from a total of 500 episodes, averaged over 100 simulation runs, was 11.79 with the doorway options (passageway state visited on 0% of episodes), compared with 9.73 with primitive actions only (passageway visited on 79% of episodes). Note that, given a certain set of assumptions, convergence on the optimal, shortest path, policy can be guaranteed in RL algorithms, including those involved in HRL. However, this is only strictly true under boundary conditions that involve extremely slow learning, due to an very slow transition from exploration to exploitation. Away from these extreme conditions, there is a marked tendency for HRL systems to “satisfice,” as illustrated in the passageway simulation.

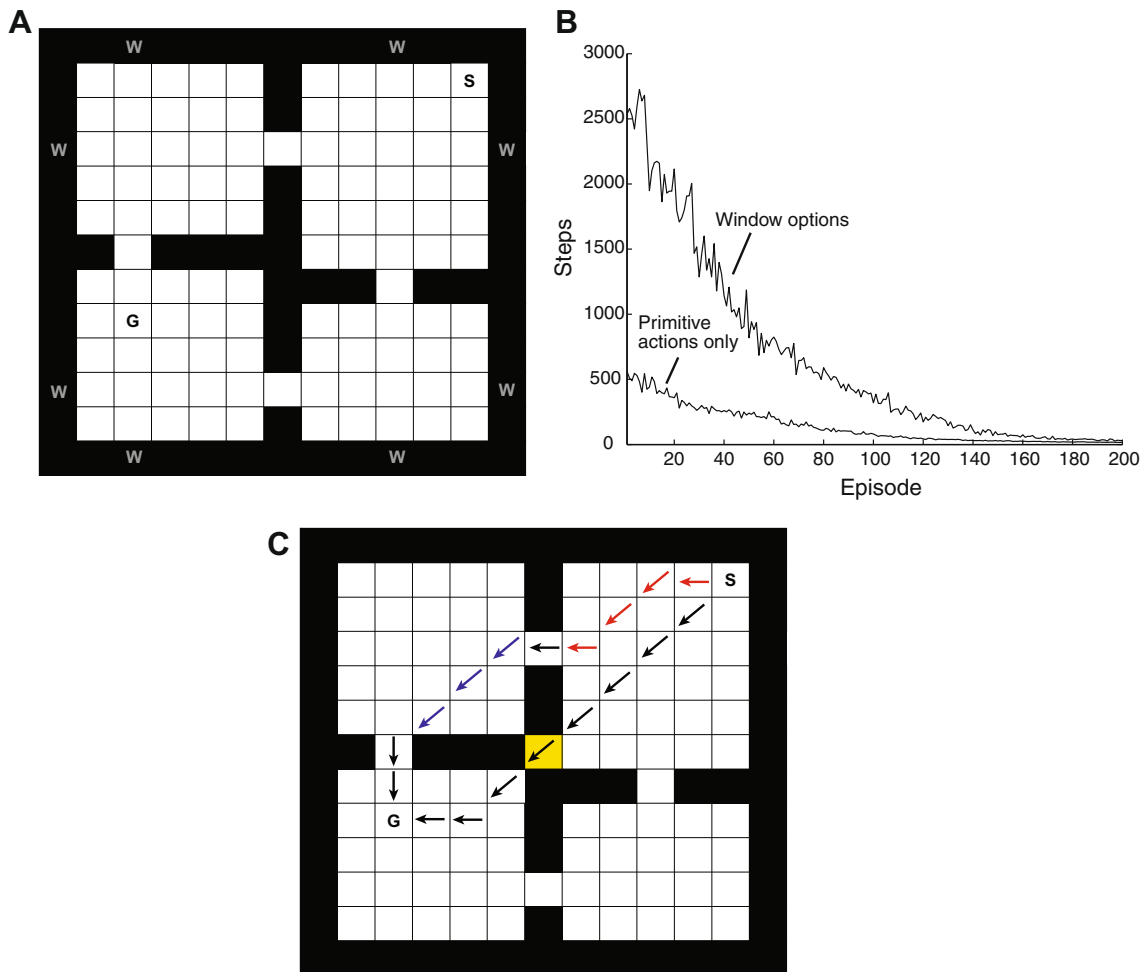


Fig. 5. (A) The rooms problem from Fig. 4, with ‘windows’ (w) defining option subgoals. (B) Learning curves for the problem illustrated in panel A. Lower data series: steps to goal over episodes with only primitive actions included (mean values over 100 simulation runs). Upper series: performance with both primitive actions and window options included. (C) Illustration of performance when a ‘shortcut’ is opened up between the upper right and lower left rooms (center tile). Lower trajectory: path to goal most frequently taken after learning with only primitive actions included. Upper trajectory: path most frequently taken after learning with both primitive actions and doorway options. Black arrows indicate primitive actions selected by the root policy. Other arrows indicate primitive actions selected by two doorway options.

positive and negative transfer in human performance indicates the relevance of this question to psychological theory, as well. With this in mind, it is of interest to consider the range of answers that have been proposed in machine learning, and their potential relations to findings from behavioral science.

4.3. The option discovery problem

One approach to the problem of discovering useful options has been to think of options as genetically specified, being shaped across generations by natural selection (Elfwing, Uchibe, & Christensen, 2007). Along these same lines, in empirical research, motor behavior has often been characterized as building upon simple, innately specified components (e.g., Bruner, 1973). In some cases extended action sequences, such as grooming sequences in rodents, have been considered to be genetically specified (Aldridge & Berridge, 1998), functioning essentially as innate options.

While evolution seems likely to play an important role in providing the building blocks for animal and human behavior, it is also clear that both animals and humans discover useful behavioral subroutines through learning (Conway & Christiansen, 2001; Fischer, 1980; Greenfield et al., 1972). One proposal from HRL for how this might be accomplished is through analysis of externally rewarded action sequences. Here, as the agent explores a particular problem, or a series of interrelated problems, it keeps a record of states or subsequences that occur relatively frequently in trajectories that culminate in reward (McGovern, 2002; Pickett & Barto, 2002; Thrun & Schwartz, 1995; see also Minton, Hayes, & Fain, 1985; Yamada & Tsujii, 1989). These states and sequences pinpoint important ‘bottlenecks’ in the problem space—such as the doors in the rooms scenario discussed above—which are good candidates to become option subgoals. On the empirical side, this proposal appears consonant with work showing that humans, even very young children, can be extremely sen-

sitive to the structure underlying repeating and systematically varying event sequences (Saffran, Aslin, & Newport, 1996), a point that extends to hierarchical structure (Saffran & Wilson, 2003).

Another HRL approach to the option discovery problem involves analyzing not trajectories through the problem space, but the problem space itself. Here, a graph is constructed to represent the relevant set of world states and the transitions that can be made among them through action. Graph partitioning methods are then used to identify states that constitute bottlenecks or access points within the graph, which are then designated as option subgoals (Mannor, Menache, Hoze, & Klein, 2004; Menache, Mannor, & Shimkin, 2002; Simsek, Wolfe, & Barto, 2005; see also Hengst, 2002; Jonsson & Barto, 2005). This set of approaches resonates with behavioral data showing that humans (including children) spontaneously generate causal representations from interactions with the world, and link these representations together into large-scale causal models (Gopnik & Schulz, 2004; Gopnik et al., 2004; Sommerville & Woodward, 2005a, 2005b). Whether such causal models are then applied toward the identification of useful subgoal states is an interesting question for empirical investigation.

Another approach within HRL takes the perspective that options can be formed during an analog of a developmental period, without the need for any externally imposed tasks. Instead of learning from extrinsically provided rewards, the agent learns from intrinsic rewards generated by built-in mechanisms that identify subgoals—states or situations that have the property that skills capable of achieving them are likely to be useful in many different future tasks (Barto, Singh, & Chentanez, 2004; Singh, Barto, & Chentanez, 2005). One example of this approach assumes that certain action outcomes are unusually salient, and that the unexpected occurrence of these outcomes during exploratory behavior triggers efforts to make them reoccur (and thus learning of options that treat these events as subgoals). More specifically, unexpected salient events are assumed to be intrinsically motivating. Singh et al. (2005) demonstrated how this mechanism can lead to the stepwise development of hierarchies of skills. The behavior of the agent in their simulations bears an intriguing similarity to children's 'circular reactions,' behavior aimed at reproducing initially inadvertent action outcomes such as turning a light on and off (Fischer & Connell, 2003; Piaget, 1936/1952). Singh et al. (2005) pointed out the unexpected occurrence of a salient event is but one way to trigger intrinsic reward, with other possibilities suggested by the psychological literature (e.g., Berlyne, 1960; White, 1959) as well as earlier studies of internal rewards in the RL literature (e.g., Kaplan & Oudeyer, 2004; Schmidhuber, 1991). Oudeyer, Kaplan, and Hafner (2007) provide an overview of much of this work.¹¹

The intrinsic motivation approach to subgoal discovery in HRL dovetails with psychological theories suggesting that human behavior is motivated by a drive toward explo-

ration or toward mastery, independent of external reward (e.g., Berlyne, 1960; Harlow, Harlow, & Meyer, 1950; Ryan & Deci, 2000; White, 1959). Moreover, the idea that unanticipated events can engage reinforcement mechanisms is also consistent with neuroscientific findings. In particular, the same midbrain dopaminergic neurons that are thought to report a temporal-difference reward prediction error also respond to salient novel stimuli (Bunzeck & Duzel, 2006; Redgrave & Gurney, 2006; Schultz, Apicella, & Ljungberg, 1993).

When option discovery is viewed as a psychological problem, other possible mechanisms for option discovery are brought to mind, which go beyond those so far considered in HRL research. For example, Soar provides a highly detailed account of subgoal generation and chunk formation, according to which subgoals, and later chunks, are established in response to problem-solving impasses (Laird et al., 1986; Lehman et al., 1996). Another still richer source of useful subgoals might be provided by the social environment. For example, empirical work with both children and adults demonstrates that human observers spontaneously infer goals and subgoals from the behavior of others (Gergely & Csibra, 2003; Meltzoff, 1995; Sommerville & Woodward, 2005a; Tenenbaum & Saxe, 2006; Woodward, Sommerville, & Guajardo, 2001). By this means, subgoals and associated action sequences could be gleaned both from deliberate demonstrations from parents, teachers, and others, and from the behavior of unwitting models (Greenfield, 1984; Yan & Fischer, 2002). Indeed, it seems natural to think of much of education and child-rearing as involving the deliberate social transmission of useful action routines. Related to this idea is the technique of shaping, whereby training is provided on low-level tasks in preparation for training on more complex tasks. In recent work, Krueger and Dayan (2008) have offered a reinforcement learning account of shaping effects, which incorporates elements, such as subtask modularity, that parallel features of HRL.

5. Neuroscientific implications

In the above, we have suggested potential bi-directional links between HRL and research on learning and behavior in humans and animals. We turn now to the potential implications of HRL for understanding neural function. To make these concrete, we will use the actor–critic formulation of HRL presented earlier. Previous work has already drawn parallels between the elements of the actor–critic framework and specific neuroanatomical structures. Situating HRL within the actor–critic framework thus facilitates the formation of hypotheses concerning how HRL might map onto functional neuroanatomy.¹²

Although accounts relating the actor–critic architecture to neural structures vary (see Joel et al. (2002), for a review), one proposal has been to identify the actor with the dorsolateral striatum (DLS), and the critic with the ven-

¹¹ These studies, directed at facilitating the learning of environmental models, are also relevant to learning of option hierarchies.

¹² For different approaches to the mapping between HRL and neuroanatomy, see De Pisapia and Goddard (2003) and Zhou and Coggins (2002, 2004).

tral striatum (VS) and the mesolimbic dopaminergic system (see, e.g., Daw, Niv, & Dayan, 2006; O'Doherty et al., 2004; Fig. 2C). Dopamine (DA), in particular, has been associated with the function of conveying reward prediction errors to both actor and critic (Barto, 1995; Montague et al., 1996; Schultz et al., 1997). In order to evaluate how HRL would modify this mapping, we will focus individually on the elements that HRL adds or modifies within the actor-critic framework, as introduced earlier. In the following two sections, we consider four key extensions, two relevant to the actor component, and two to the critic.

5.1. The actor in HRL: relation to prefrontal cortex

5.1.1. Extension 1: support structure for temporally abstract actions

Under HRL, in addition to primitive actions, the actor must build in representations that identify specific temporally abstract actions or options. Using these, the actor must be able to keep track of which option is currently selected and in control of behavior.

5.1.2. Potential neural correlates

This first extension to the actor-critic framework calls to mind functions commonly ascribed to the dorsolateral prefrontal cortex (DLPFC). The DLPFC has long been considered to house representations that guide temporally integrated, goal-directed behavior (Fuster, 1997, 2004; Grafman, 2002; Petrides, 1995; Shallice & Burgess, 1991; Wood & Grafman, 2003). Recent work has refined this idea by demonstrating that DLPFC neurons play a direct role in representing *task sets*. Here, a single pattern of DLPFC activation serves to represent an entire mapping from stimuli to responses, i.e., a policy (Asaad, Rainer, & Miller, 2000; Bunge, 2004; Hoshi, Shima, & Tanji, 1998; Johnston & Everling, 2006; Rougier, Noell, Braver, Cohen, & O'Reilly, 2005; Shimamura, 2000; Wallis, Anderson, & Miller, 2001; White, 1999). According to the guided activation theory proposed by Miller and Cohen (2001), prefrontal representations do not implement policies directly, but instead select among stimulus-response pathways implemented outside the prefrontal cortex. This division of labor fits well with the distinction in HRL between an option's identifier and the policy with which it is associated (Fig. 6).

In addition to the DLPFC, there is evidence that other frontal areas may also carry representations of task set, including pre-supplementary motor area (pre-SMA; Rushworth, Walton, Kennerley, & Bannerman, 2004) and premotor cortex (PMC; Muhammad, Wallis, & Miller, 2006; Wallis & Miller, 2003). Furthermore, like options in HRL, neurons in several frontal areas including DLPFC, pre-SMA and supplementary motor area (SMA) have been shown to code for particular sequences of low-level actions (Averbeck & Lee, 2007; Bor, Duncan, Wiseman, & Owen, 2003; Shima, Isoda, Mushiake, & Tanji, 2007; Shima & Tanji, 2000). Research on frontal cortex also accords well with the stipulation in HRL that temporally abstract actions may organize into hierarchies, with the policy for one option (say, an option for making coffee) calling other, lower-level options (say, options for adding sugar or cream). This fits with numerous accounts suggesting that the frontal cortex

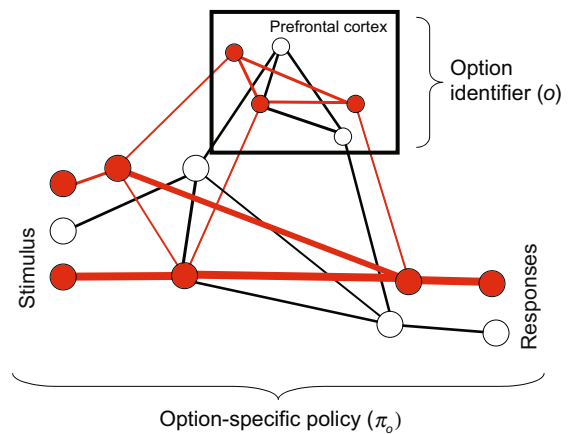


Fig. 6. Illustration of the role of the prefrontal cortex, as postulated by guided activation theory (Miller & Cohen, 2001). Patterns of activation in prefrontal cortex (filled elements in the boxed region) effectively select among stimulus-response pathways lying elsewhere in the brain (lower area). Here, representations within prefrontal cortex correspond to option identifiers in HRL, while the stimulus-response pathways selected correspond to option-specific policies. Figure adapted from Miller and Cohen (2001, Reprinted, with permission, from the Annual Review of Neuroscience, Volume 24 (c) 2001 by Annual Reviews <<http://www.annualreviews.org/>> www.annualreviews.org).

serves to represent action at multiple, nested levels of temporal structure (Grafman, 2002; Sirigu et al., 1995; Wood & Grafman, 2003; Zalla, Pradat-Diehl, & Sirigu, 2003), possibly in such a way that higher levels of structure are represented more anteriorly (Botvinick, 2008; Fuster, 2001, 2004; Haruno & Kawato, 2006; Koehlin et al., 2003).

5.1.3. Extension 2: option-specific policies

In addition to its default, top-level policy, the actor in HRL must implement option-specific policies. Thus, the actor must carry a separate set of action strengths for each option.

5.1.4. Potential neural correlates

As noted earlier, it has been typical to draw a connection from the policy in standard RL to the DLS. For the DLS to implement the option-specific policies found in HRL, it would need to receive input from cortical regions representing options. It is thus relevant that such regions as the DLPFC, SMA, pre-SMA and PMC—areas potentially representing options—all project heavily to the DLS (Alexander, DeLong, & Strick, 1986; Parent & Hazrati, 1995). Frank, O'Reilly and colleagues (Frank & Claus, 2006; O'Reilly & Frank, 2006; Rougier, Noell, Braver, Cohen, & O'Reilly, 2005) have put forth detailed computational models that show how frontal inputs to the striatum could switch among different stimulus-response pathways. Here, as in guided activation theory, temporally abstract action representations in frontal cortex select among alternative (i.e., option-specific) policies.

In order to support option-specific policies, the DLS would need to integrate information about the currently controlling option with information about the current environmental state, as is indicated by the arrows converging on the policy module in Fig. 2B. This is consistent with neurophysiological data showing that some DLS neurons

respond to stimuli in a way that varies with task context (Ravel, Sardo, Legallet, & Apicella, 2006; see also Salinas, 2004). Other studies have shown that *action* representations within the DLS can also be task-dependent. For example, Aldridge and Berridge (1998) reported that, in rats, different DLS neurons fired in conjunction with simple grooming movements depending on whether those actions were performed in isolation or as part of a grooming sequence (see also Aldridge, Berridge, & Rosen, 2004; Graybiel, 1995, 1998; Lee, Seitz, & Assad, 2006). This is consistent with the idea that option-specific policies (action strengths) might be implemented in the DLS, since this would imply that a particular motor behavior, when performed in different task contexts, would be selected via different neural pathways.

Recall that, within HRL, policies are responsible for selecting not only primitive actions, but also for selecting options. Translated into neural terms, this would require the DLS to participate in the selection of options. This is consistent with data from Muhammad et al. (2006), who observed striatal activation that varied with task context (see also Graybiel, 1998). It is also consistent with the fact that the DLS projects heavily, via thalamic relays, to all of the frontal regions linked above with a role in representing options (Alexander et al., 1986; Middleton & Strick, 2002).

Unlike the selection of primitive actions, the selection of options in HRL involves initiation, maintenance and termination phases. At the neural level, the maintenance phase would be naturally supported within DLPFC, which has been extensively implicated in working memory function (Courtney et al., 2007; D'Esposito, 2007; Postle, 2006). With regard to initiation and termination, it is intriguing that phasic activity has been observed, both within the DLS and in several areas of frontal cortex, at the boundaries of temporally extended action sequences (Fujii & Graybiel, 2003; Morris, Arkadir, Nevet, Vaadia, & Bergman, 2004; Zacks et al., 2001). Since these boundaries correspond to points where new options would be selected, boundary-aligned activity in the DLS and frontal cortex is also consistent with a proposed role of the DLS in gating information into prefrontal working memory circuits (O'Reilly & Frank, 2006; Rougier, Noell, Braver, Cohen, & O'Reilly, 2005).

5.2. The critic in HRL: relation to orbitofrontal cortex

As noted earlier, HRL also requires two key extensions to the critic component of the actor–critic architecture.

5.2.1. Extension 3: option-specific value functions

Under HRL, in addition to its top-level state-value function, the critic must also maintain a set of option-specific value functions. This is due to the fact that the value function indicates how well things are expected to go following arrival at a given state, which obviously depends on which actions the agent will select. Under HRL, the option that is currently in control of behavior determines action selection, and also determines which actions will yield pseudo-reward. Thus, whenever an option is guiding behavior, the value attached to a state must take the identity of that option into account.

5.2.2. Potential neural correlates

If there is a neural structure that computes something like option-specific state values, this structure would be expected to communicate closely with the VS, the region typically identified with the locus of state or state-action values in RL. However, the structure would also be expected to receive inputs from the portions of frontal cortex that we have identified as representing options. One brain region that meets both of these criteria is the orbitofrontal cortex (OFC), an area that has strong connections with both VS and DLPFC (Alexander, Crutcher, & DeLong, 1990; Rolls, 2004). The idea that the OFC might participate in computing option-specific state values also fits well with the behavior of individual neurons within this cortical region. OFC neurons have been extensively implicated in representing the reward value associated with environmental states (Rolls, 2004; Schultz, Tremblay, & Hollerman, 2000). However, other data suggests that OFC neurons can also be sensitive to shifts in response policy or task set (e.g., O'Doherty, Critchley, Deichmann, & Dolan, 2003). Critically, Schoenbaum, Chiba, and Gallagher (1999) observed that OFC representations of event value changed in parallel with shifts in strategy, a finding that fits precisely with the idea that the OFC might represent option-specific state values.

5.2.3. Extension 4: temporal scope of the prediction error

Moving from RL to HRL brings about an important alteration in the way that the prediction error is computed. Specifically, it changes the scope of the events that the prediction error addresses. In standard RL, the prediction error indicates whether things went better or worse than expected since the immediately preceding time-step. HRL, in addition, evaluates at the completion of an option whether things have gone better or worse than expected since the initiation of that option (see Fig. 3). Thus, unlike standard RL, the prediction errors associated with options in HRL are framed around temporally extended events. Formally speaking, the HRL setting is no longer a Markov decision process, but rather a semi-Markov decision process (SMDP).

5.2.4. Potential neural correlates

This aspect of HRL resonates, once again, with data from the OFC. Note that, in order to evaluate whether things went better or worse than expected over the course of an entire option, the critic needs access, when an option terminates, to the reward prediction it made when the option was initially selected. This is consistent with the finding that within OFC, unlike some other areas, reward-predictive activity tends to be sustained, spanning temporally extended segments of task structure (Schultz et al., 2000). Another relevant finding is that the response of OFC neurons to the receipt of primary rewards varies depending on the wait-time leading up to the reward (Roesch, Taylor, & Schoenbaum, 2006; see Appendix, Eq. (7)). This suggests, again, that the OFC interprets value within the context of temporally extended segments of behavior.

The widened scope of the prediction error computation in HRL also resonates with work on midbrain DA function. In particular, Daw, Courville, and Touretzky (2003) suggested, based on midbrain responses to delayed rewards,

that dopaminergic function is driven by representations that divide event sequences into temporally extended segments. In articulating this account, Daw et al. (2003) provided a formal analysis of DA function that draws on precisely the same principles of temporal abstraction that also provide the foundation for HRL, namely an SMDP framework.

In further examining the potential links between DA and HRL, it may be useful to consider recent work by O'Reilly and Frank (2006), which shows through computational modeling how DA might support learning in working memory circuits, supporting the performance of hierarchically organized, temporally extended tasks. This research addresses issues somewhat different from those that are central to HRL, focusing in particular upon tasks that require preservation of information conveyed by transient cues (a case treated in machine learning under the rubric of partially observable Markov decision problems). However, O'Reilly and colleagues have also begun to explore the application of similar mechanisms to the learning of abstract task representations (Rougier et al., 2005). One interesting aspect of this latter work is its focus on cases where task-appropriate behavior can be acquired through attending selectively to particular stimulus dimensions (e.g., color or shape). This connects with some work in HRL, where the use of option-specific state representations have been explored (see, e.g., Dietterich, 2000; Jonsson & Barto, 2001). Characterizing further the relationship between this approach within HRL and the computational work by Rougier and colleagues is an inviting area for further analysis.

6. Discussion

We have shown that recently developed HRL techniques have much in common with psychological accounts of hierarchically organized behavior. Furthermore, through a new actor–critic implementation of HRL, we have suggested several points of contact between HRL and the neural substrates of decision making and hierarchical control. Before summing up, we briefly consider the relation of HRL to two further topics that have been at the focus of recent work on the control of action, and we enumerate some directions for further research.

6.1. Dual modes of action control

Work on animal and human behavior suggests that instrumental actions arise from two modes of control, one built on established stimulus–response links or ‘habits,’ and the other on prospective planning (Balleine & Dickinson, 1998). Daw et al. (2005) have mapped these modes of control onto RL constructs, characterizing the former as relying on cached action values or strengths and model-free RL, and the latter as looking ahead based on an internal model relating actions to their likely effects, that is, model-based RL. Here we have cast HRL in terms of the cache-based system, both because this is most representative of existing work on HRL and because the principles of model-based search have not yet been as fully explored, either at the computational level or in terms of

neural correlates. However, it is straightforward to incorporate temporal abstraction into model-based, prospective control. This is accomplished by assuming that each option is associated with an *option model*, a knowledge structure indicating the ultimate outcomes likely to result from selecting the option, the reward or cost likely to be accrued during its execution, and the amount of time this execution is likely to take (see Sutton et al., 1999). Equipped with models of this kind, the agent can use them to look ahead, evaluating potential courses of action. Importantly, the search process can now ‘skip over’ potentially large sequences of primitive actions, effectively reducing the size of the search tree (Fig. 1C; Hayes-Roth & Hayes-Roth, 1979; Kambhampati, Mali, & Srivastava, 1998; Marthi, Russell, & Wolfe, 2007). This kind of saltatory search process seems to fit well with everyday planning, which introspectively seems to operate at the level of temporally abstract actions (‘Perhaps I should buy one of those new cell phones... Well, that would cost me a few hundred dollars... But if I bought one, I could use it to check my email...’). The idea of action models, in general, also fits well with work on motor control (e.g., Wolpert & Flanagan, 2001), which strongly suggests the involvement of predictive models in the guidance of bodily movements. Because option models encode the consequences of interventions, it is interesting to note that recent neuroimaging work has mapped representations of action outcome information in part to prefrontal cortex (Hamilton & Grafton, 2008), a region whose potential links with HRL we have already considered.

6.2. Strict versus quasi-hierarchical structure

Although human behavior, like behavior in HRL systems, is often hierarchically structured, there are also aspects of human behavior that resist a strictly hierarchical account (Botvinick, 2007, 2008; Botvinick & Plaut, 2002, 2004, 2006). For example, naturalistic tasks exhibit a great deal of overlap or shared structure (Schank & Abelson, 1977), a point that is reflected in the errors or slips that occur in the performance of such tasks (Reason, 1992). Shared structure raises a problem because temporal abstractions have only a limited ability to exploit detailed patterns of overlap among tasks. Thus, using options (as they have so far been defined), it would be difficult to capture the overlap among tasks such as spreading jam on bread, spreading mustard on a hotdog, and spreading icing on a cake. Furthermore, execution of subtasks in everyday behavior is highly context-sensitive, that is, the way in which a subtask is executed can depend on the larger task context in which it occurs (Agre, 1988). Context sensitivity raises the problem that different levels within a task hierarchy are no longer independent. For example, the subtask of picking up a pencil cannot be represented as a self-contained unit if the details of its execution (e.g., the rotation of the hand) depend on whether one is going to use the pencil to write or to erase (see Ansuini, Santello, Massaccesi, & Castiello, 2006). Significantly, related tensions between hierarchical compositionality and context-sensitivity have also been noted in work on HRL (Dietterich, 2000).

Botvinick and Plaut (2002, 2004, 2006) proposed a computational model of routine sequential behavior that is sensitive to hierarchical task structure, but which also accommodates context-dependent subtask performance and overlap between tasks. That model, like the HRL model we have presented here, displays transfer effects when faced with new problems (Botvinick & Plaut, 2002). Furthermore, Ruh (2007) has demonstrated that the Botvinick and Plaut (2004) model can acquire target behaviors through RL. Understanding the relationship between this computational approach and HRL is an interesting challenge for further investigation.

6.3. Directions for further research

The idea that HRL algorithms may be relevant to understanding brain function and behavior gives rise to a wide range of questions for experimental research. To begin with, almost all of the empirical parallels we have traced out in the present article call for further experimental scrutiny. For example, although there is compelling evidence that dorsolateral PFC neurons represent task context, there is as of yet only indirect evidence (e.g., Badre, 2008; Zacks et al., 2001) to support the idea that PFC neurons also code discretely for subtask segments in hierarchically structured tasks, as HRL would seem to require. Similarly, although there are established aspects of dopaminergic function that resonate with predictions from HRL, HRL also gives rise to predictions that would need to be tested through new experiments. In particular, the version of HRL we have focused on here predicts phasic dopaminergic discharge at subtask boundaries, scaling with the magnitude of option-level prediction errors.

In considering the coding of rewards, one particularly interesting question is whether there might exist a neural correlate of pseudo-reward. As detailed earlier, the options framework, like at least one other influential version of HRL (MAXQ, Dietterich, 2000), associates a separate reward function with each individual subtask representation. This pseudo-reward function is critical in shaping subtask-specific policies, directing action toward desirable subtask outcomes, i.e., subgoals. An obvious question is whether neural structures that are responsive to exogenous reward also respond to the attainment of subgoal states during the performance of hierarchically structured tasks.

Having suggested this possibility, it is important to note that in the case of pseudo-reward, the framing of specific predictions requires consideration of differences between HRL implementations. Indeed, there exist versions of HRL, such as the HAM framework (Parr & Russell, 1998), that do not involve pseudo-reward at all, relying instead on a combination of ordinary, exogenous reward and fixed constraints to shape subtask policies. Another point on which HRL implementations differ, which may have implications for experimental predictions, is in the representation of the value function. Whereas the options framework, as we have discussed, maintains option-specific value functions, some other frameworks, including MAXQ (Andre & Russell, 2001, 2002; Dietterich, 2000), decompose these value functions into separate components, giving rise to slightly different learning algorithms. This difference may lead different

HRL frameworks to make non-equivalent fine-grained predictions. For example, frameworks may vary in the prediction errors they would compute in a given domain, giving rise to divergent predictions concerning dopaminergic function. In sum, just as with non-hierarchical RL algorithms, any detailed investigation of HRL as a framework for interpreting neural function and behavior will require some attention to algorithmic detail.

7. Conclusion

Computational RL has proved extremely useful to research on behavior and brain function. Our aim here has been to explore whether HRL might prove similarly applicable. An initial motivation for considering this question derives from the fact that HRL addresses an inherent limitation of RL, the scaling problem, which would clearly be of relevance to any organism relying on RL-like learning mechanisms. Implementing HRL along the lines of the actor–critic framework, thereby bringing it into alignment with existing mappings between RL and neuroscience, reveals direct parallels between components of HRL and specific functional neuroanatomic structures, including the DLPFC and OFC. HRL suggests new ways of interpreting neural activity in these as well as several other regions. HRL also resonates strongly with issues in psychology, in particular with work on task representation and the control of hierarchically structured behavior, adding to these a unifying normative perspective. Among the most important implications of HRL is the way in which it highlights the option discovery problem. Here, and on many other fronts, HRL appears to offer a potentially useful set of tools for further investigating the computational and neural basis of hierarchical structured behavior.

Acknowledgements

The present work was completed with support from the National Institute of Mental Health, Grant No. P50 MH062196 (M.M.B.), the Human Frontiers Science Program (Y.N.), and from the National Science Foundation, Grant No. CCF-0432143 (A.C.B.). Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding agencies. The authors thank Carlos Brody, Jonathan Cohen, Scott Kuindersma, Ken Norman, Randy O'Reilly, Geoff Schoenbaum, Asvin Shah, Ozgur Simsek, Andrew Stout, Chris Vigorito, and Pippin Wolfe for useful comments on the work reported.

Appendix A

We present here the details of our HRL implementation and the simulations briefly described in the main text. For clarity, we begin by describing our implementation of non-hierarchical RL, which was used in the simulations including only primitive actions. This will then be extended, in the next section, to the hierarchical case. All simulations were run using Matlab (The Mathworks, Na-

tick, MA). Code is available for download at www.princeton.edu/~matthewb.

A.1. Basic actor–critic implementation

A.1.1. Task and representations

Following the standard RL approach (see Sutton & Barto, 1998), tasks were represented by four elements: a set of states S , a set of actions A , a reward function R assigning a real-valued number to every state transition, and a transition function T giving a new state for each pairing of a state with an action. In our simulations, S contained the set of location tiles in the layout depicted in Fig. 4A; A contained eight single-step movements, following the principle compass directions; R yielded a reward of 100 on transitions to the goal state indicated with a G in Fig. 4A, otherwise zero; and T was deterministic. All actions were available in every state, and actions yielded no change in state if a move into a wall was attempted. Our choice to use deterministic actions was for simplicity of exposition, and does not reflect a limitation of either the RL or HRL framework.

A.1.2. Architecture

The basic RL agent comprised actor and critic components. The actor maintained a set (matrix) of real-valued strengths (W) for each action in each state. The critic maintained a vector V of values, attaching a real number to each state.

A.1.3. Training

At the outset of training, action strengths and state values were initialized to zero; the state was initialized to the start location indicated in Fig. 4A; and a time index t was initialized at zero. On each step of processing, t , an action was selected probabilistically according to the softmax equation:

$$P(a) = \frac{e^{W(s_t, a)/\tau}}{\sum_{a' \in A} e^{W(s_t, a')/\tau}} \quad (1)$$

where $P(a)$ is the probability of selecting action a at step t ; $W(s_t, a)$ is the weight for action a in the current state; and τ is a temperature parameter controlling the tendency toward exploration in action selection (10 in our simulations). The next state (s_{t+1}) was then determined based on the transition function T , and the reward for the transition (r_{t+1}) based on R . Using these, the temporal-difference (TD) prediction error (δ) was computed as

$$\delta = r_{t+1} + \gamma V(s_{t+1}) - V(s_t) \quad (2)$$

where γ is a discount factor (0.9 in our simulations). The TD prediction error was then used to update both the value function and the strength for the action just completed:

$$V(s_t) \leftarrow V(s_t) + \alpha_c \delta \quad (3)$$

$$W(s_t, a) \leftarrow W(s_t, a) + \alpha_a \delta \quad (4)$$

The learning rate parameters α_c and α_a were set to 0.2 and 0.1, respectively. Following these updates, t was incremented and a new action was selected. The cycle was repeated until the goal state was reached, at which point

the agent was returned to the start state, t was reinitialized, and another episode was run.

A.2. HRL implementation

Our implementation of HRL was based on the options framework described by Sutton et al. (1999), but adapted to the actor–critic framework.

A.2.1. Task and representations

The set of available actions was expanded to include options in addition to primitive actions. Each option was associated with (1) an initiation set, indicating the states where the option could be selected; (2) a termination function, returning the probability of terminating the option in each state; and (3) a set of option-specific strengths W_o , containing one weight for each action (primitive or abstract) at each state.

For the four-rooms simulations, two options could be initiated in each room, each terminating deterministically at one of the room's two doors. Each option also had a pseudo-reward function, yielding a pseudo-reward of 100 at the option's termination state. For simplicity, each option was associated with strengths only for primitive actions (i.e., not for other options). That is, option policies were only permitted to select primitive actions. As indicated in the main text, options are ordinarily permitted to select other options. This more general arrangement is compatible with the implementation described here.

A.2.2. Architecture

In addition to the option-specific strengths just mentioned, the actor maintained a 'root' set of strengths, used for action selection when no option was currently active. The critic maintained a root-level value function plus a set of option-specific value functions V_o .

A.2.3. Training

Since primitive actions can be thought of as single-step options, we shall henceforth refer to primitive actions as 'primitive options' and temporally abstract actions as 'abstract options,' using the term 'option' to refer to both at once. The model was initialized as before, with all option strengths and state values initialized to zero. On each successive step, an option o was selected according to

$$P(o) = \frac{e^{W_{o_{\text{ctrl}}}(s_t, o)/\tau}}{\sum_{o' \in O} e^{W_{o'_{\text{ctrl}}}(s_t, o')/\tau}} \quad (5)$$

where O is the set of available options, including primitive options; o_{ctrl} is the option currently in control of behavior (if any); and $W_{o_{\text{ctrl}}}(s_t, o)$ is the option-specific—i.e., o_{ctrl} -specific—strength for option o (or the root strength for o in the case where no option is currently in control). Following identification of the next state and of the reward (including pseudo-reward) yielded by the transition, the prediction error was calculated for all *terminating* options, including primitive options, as

$$\delta = r_{\text{cum}} + \gamma^{t_{\text{tot}}} V_{o_{\text{ctrl}}}(s_{t+1}) - V_{o_{\text{ctrl}}}(s_{\text{init}}) \quad (6)$$

where t_{tot} is the number of time-steps elapsed since the relevant option was selected (one for primitive actions); $s_{t_{\text{init}}}$ is the state in which the option was selected; o_{ctrl} is the option whose policy selected the option that is now terminating (or the root value function if the terminating option was selected by the root policy); and r_{cum} is the cumulative discounted reward for the duration of the option:

$$r_{\text{cum}} = \sum_{i=1}^{t_{\text{tot}}} \gamma^{i-1} r_{t_{\text{init}}+i} \quad (7)$$

Note that $r_{t_{\text{init}}+i}$ incorporated pseudo-reward only if $s_{t_{\text{init}}+i}$ was a subgoal state for o_{ctrl} . Thus, pseudo-reward was used to compute prediction errors ‘within’ an option, i.e., when updating the option’s policy, but not ‘outside’ the option, at the next level up. It should also be remarked that, at the termination of non-primitive options, two TD prediction errors were computed, one for the last primitive action selected under the option and one for the option itself (see Fig. 3).

Following calculation of each δ , value functions and option strengths were updated:

$$V_{o_{\text{ctrl}}}(s_{t_{\text{init}}}) \leftarrow V_{o_{\text{ctrl}}}(s_{t_{\text{init}}}) + \alpha_C \delta \quad (8)$$

$$W_{o_{\text{ctrl}}}(s_{t_{\text{init}}}, o) \leftarrow W_{o_{\text{ctrl}}}(s_{t_{\text{init}}}, o) + \alpha_A \delta \quad (9)$$

The time index was then incremented and a new option/action selected, with the entire cycle continuing until the top-level goal was reached.

In our simulations, the model was first pre-trained for a total of 50,000 time-steps without termination or reward delivery at G. This allowed option-specific action strengths and values to develop, but did not lead to any change in strengths or values at the root level. Thus, action selection at the top-level was random during this phase of training. In order to obtain the data displayed in Fig. 4C, for clarity of illustration, training with pseudo-reward only was conducted with a small learning rate ($\alpha_A = 0.01$, $\alpha_C = 0.1$), reinitializing to a random state whenever the relevant option reached its subgoal.

References

- Agre, P. E. (1988). *The dynamic structure of everyday life* (Tech. Rep. No. 1085). Cambridge, MA: Massachusetts Institute of Technology, Artificial Intelligence Laboratory.
- Aldridge, W. J., & Berridge, K. C. (1998). Coding of serial order by neostriatal neurons: A “natural action” approach to movement sequence. *Journal of Neuroscience*, 18, 2777–2787.
- Aldridge, J. W., Berridge, K. C., & Rosen, A. R. (2004). Basal ganglia neural mechanisms of natural movement sequences. *Canadian Journal of Physiology and Pharmacology*, 82, 732–739.
- Alexander, G. E., Crutcher, M. D., & DeLong, M. R. (1990). Basal ganglia-thalamocortical circuits: Parallel substrates for motor, oculomotor, “prefrontal” and “limbic” functions. *Progress in Brain Research*, 85, 119–146.
- Alexander, G. E., DeLong, M. R., & Strick, P. L. (1986). Parallel organization of functionally segregated circuits linking basal ganglia and cortex. *Annual Review of Neuroscience*, 9, 357–381.
- Allport, A., & Wylie, G. (2000). Task-switching, stimulus-response bindings and negative priming. In S. Monsell & J. Driver (Eds.), *Control of cognitive processes: Attention and performance XVIII*. Cambridge, MA: MIT Press.
- Anderson, J. R. (2004). An integrated theory of mind. *Psychological Review*, 111, 1036–1060.
- Andre, D., & Russell, S. J. (2001). Programmable reinforcement learning agents. *Advances in Neural Information Processing Systems*, 13, 1019–1025.
- Andre, D., & Russell, S. J. (2002). State abstraction for programmable reinforcement learning agents. *Paper presented at the proceedings of the 18th national conference on artificial intelligence*.
- Ansuini, C., Santello, M., Massaccesi, S., & Castiello, U. (2006). Effects of end-goal on hand shaping. *Journal of Neurophysiology*, 95, 2456–2465.
- Arbib, M. A. (1985). Schemas for the temporal organization of behaviour. *Human Neurobiology*, 4, 63–72.
- Asaad, W. F., Rainer, G., & Miller, E. K. (2000). Task-specific neural activity in the primate prefrontal cortex. *Journal of Neurophysiology*, 84, 451–459.
- Averbeck, B. B., & Lee, D. (2007). Prefrontal neural correlates of memory for sequences. *Journal of Neuroscience*, 27, 2204–2211.
- Badre, D. (2008). Cognitive control, hierarchy, and the rostro-caudal organization of the frontal lobes. *Trends in Cognitive Sciences*, 12, 193–200.
- Balleine, B. W., & Dickinson, A. (1998). Goal-directed instrumental action: Contingency and incentive learning and their cortical substrates. *Neuropharmacology*, 37, 407–419.
- 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: Theory and Applications*, 13, 343–379.
- Barto, A. G., Singh, S., & Chentanez, N. (2004). Intrinsically motivated learning of hierarchical collections of skills. In *Proceedings of the 3rd international conference on development and learning (ICDL 2004)*.
- Barto, A. G., & Sutton, R. S. (1981). Toward a modern theory of adaptive networks: Expectation and prediction. *Psychological Review*, 88, 135–170.
- Barto, A. G., Sutton, R. S., & Anderson, C. W. (1983). Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE Transactions on Systems, Man and Cybernetics*, 13, 834–846.
- Berlyne, D. E. (1960). *Conflict, arousal and curiosity*. New York: McGraw-Hill.
- Bhatnagara, S., & Panigrahi, J. R. (2006). Actor-critic algorithms for hierarchical Markov decision processes. *Automatica*, 42, 637–644.
- Bor, D., Duncan, J., Wiseman, R. J., & Owen, A. M. (2003). Encoding strategies dissociate prefrontal activity from working memory demand. *Neuron*, 37, 361–367.
- Botvinick, M. M. (2007). Multilevel structure in behaviour and the brain: A model of Fuster’s hierarchy. *Philosophical Transactions of the Royal Society (London), Series B*, 362, 1615–1626.
- Botvinick, M. M. (2008). Hierarchical models of behavior and prefrontal function. *Trends in Cognitive Sciences*, 12, 201–208.
- Botvinick, M., & Plaut, D. C. (2002). Representing task context: Proposals based on a connectionist model of action. *Psychological Research*, 66(4), 298–311.
- Botvinick, M., & 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., & Plaut, D. C. (2006). Such stuff as habits are made on: A reply to Cooper and Shallice (2006). *Psychological Review*, under review.
- Bruner, J. (1973). Organization of early skilled action. *Child Development*, 44, 1–11.
- Bunge, S. A. (2004). How we use rules to select actions: A review of evidence from cognitive neuroscience. *Cognitive, Affective and Behavioral Neuroscience*, 4, 564–579.
- Bunzeck, N., & Duzel, E. (2006). Absolute coding of stimulus novelty in the human substantia nigra/VTA. *Neuron*, 51, 369–379.
- Cohen, J. D., Braver, T. S., & O’Reilly, R. C. (1996). A computational approach to prefrontal cortex, cognitive control and schizophrenia: Recent developments and current challenges. *Philosophical Transactions of the Royal Society (London), Series B*, 351, 1515–1527.
- Cohen, J. D., Dunbar, K., & McClelland, J. L. (1990). On the control of automatic processes: A parallel distributed processing account of the Stroop effect. *Psychological Review*, 97, 332–361.
- Conway, C. M., & Christiansen, M. H. (2001). Sequential learning in non-human primates. *Trends in Cognitive Sciences*, 5, 539–546.
- Cooper, R., & Shallice, T. (2000). Contention scheduling and the control of routine activities. *Cognitive Neuropsychology*, 17, 297–338.
- Courtney, S. M., Roth, J. K., & Sala, J. B. (2007). A hierarchical biased-competition model of domain-dependent working memory maintenance and executive control. In N. Osaka, R. Logie, & M. D’Esposito (Eds.), *Working memory: Behavioural and neural correlates* (pp. 369–383). Oxford: Oxford University Press.
- D’Esposito, M. (2007). From cognitive to neural models of working memory. *Philosophical Transactions of the Royal Society (London), Series B*, 362, 761–772.

- Daw, N. D., Courville, A. C., & Touretzky, D. S. (2003). *Timing and partial observability in the dopamine system. Advances in neural information processing systems* (Vol. 15). Cambridge, MA: MIT Press (pp. 99–106).
- Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and striatal systems for behavioral control. *Nature Neuroscience*, 8, 1704–1711.
- Daw, N. D., Niv, Y., & Dayan, P. (2006). Actions, policies, values and the basal ganglia. In E. Bezdard (Ed.), *Recent breakthroughs in basal ganglia research*. New York: Nova Science Publishers.
- De Pisapia, N., & Goddard, N. H. (2003). A neural model of frontostriatal interactions for behavioral planning and action chunking. *Neurocomputing*, 489–495.
- Dehaene, S., & Changeux, J.-P. (1997). A hierarchical neuronal network for planning behavior. *Proceedings of the National Academy of Sciences*, 94, 13293–13298.
- Dell, G. S., Berger, L. K., & Svec, W. R. (1997). Language production and serial order. *Psychological Review*, 104, 123–147.
- Dieterich, T. G. (1998). The MAXQ method for hierarchical reinforcement learning. In *Proceedings of the international conference on machine learning*.
- Dieterich, T. G. (2000). Hierarchical reinforcement learning with the MAXQ value function decomposition. *Journal of Artificial Intelligence Research*, 13, 227–303.
- Elfwing, S., Uchibe, K., & Christensen, H. I. (2007). Evolutionary development of hierarchical learning structures. *IEEE Transactions on Evolutionary Computations*, 11, 249–264.
- Estes, W. K. (1972). An associative basis for coding and organization in memory. In A. W. Melton & E. Martin (Eds.), *Coding processes in human memory* (pp. 161–190). Washington, DC: V.H. Winston & Sons.
- Fischer, K. W. (1980). A theory of cognitive development: The control and construction of hierarchies of skills. *Psychological Review*, 87, 477–531.
- Fischer, K. W., & Connell, M. W. (2003). Two motivational systems that shape development: Epistemic and self-organizing. *British Journal of Educational Psychology: Monograph Series II*, 2, 103–123.
- Frank, M. J., & Claus, E. D. (2006). Anatomy of a decision: Striato-orbitofrontal interactions in reinforcement learning, decision making, and reversal. *Psychological Review*, 113, 300–326.
- Fujii, N., & Graybiel, A. M. (2003). Representation of action sequence boundaries by macaque prefrontal cortical neurons. *Science*, 301, 1246–1249.
- Fuster, J. M. (1997). *The prefrontal cortex: Anatomy, physiology, and neuropsychology of the frontal lobe*. Philadelphia, PA: Lippincott-Raven.
- Fuster, J. M. (2001). The prefrontal cortex—An update: Time is of the essence. *Neuron*, 30, 319–333.
- Fuster, J. M. (2004). Upper processing stages of the perception–action cycle. *Trends in Cognitive Sciences*, 8, 143–145.
- Gergely, G., & Csibra, G. (2003). Teleological reasoning in infancy: The naive theory of rational action. *Trends in Cognitive Sciences*, 7, 287–292.
- Gopnik, A., Glymour, C., Sobel, D., Schulz, T., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review*, 111, 1–31.
- Gopnik, A., & Schulz, L. (2004). Mechanisms of theory formation in young children. *Trends in Cognitive Sciences*, 8, 371–377.
- Grafman, J. (2002). The human prefrontal cortex has evolved to represent components of structured event complexes. In J. Grafman (Ed.), *Handbook of neuropsychology*. Amsterdam: Elsevier.
- Graybiel, A. M. (1995). Building action repertoires: Memory and learning functions of the basal ganglia. *Current Opinion in Neurobiology*, 5, 733–741.
- Graybiel, A. M. (1998). The basal ganglia and chunking of action repertoires. *Neurobiology of Learning and Memory*, 70, 119–136.
- Greenfield, P. M. (1984). A theory of the teacher in the learning activities of everyday life. In B. Rogoff & J. Lave (Eds.), *Everyday cognition: Its development in social context* (pp. 117–138). Cambridge, MA: Harvard University Press.
- Greenfield, P. M., Nelson, K., & Saltzman, E. (1972). The development of rulebound strategies for manipulating seriated cups: A parallel between action and grammar. *Cognitive Psychology*, 3, 291–310.
- Greenfield, P. M., & Schneider, L. (1977). Building a tree structure: The development of hierarchical complexity and interrupted strategies in children's construction activity. *Developmental Psychology*, 13, 299–313.
- Grossberg, S. (1986). The adaptive self-organization of serial order in behavior: Speech, language, and motor control. In E. C. Schwab & H. C. Nusbaum (Eds.), *Pattern recognition by humans and machines, vol. 1: Speech perception* (pp. 187–294). New York: Academic Press.
- Hamilton, A. F. d. C., & Grafton, S. T. (2008). Action outcomes are represented in human inferior frontoparietal cortex. *Cerebral Cortex*, 18, 1160–1168.
- Harlow, H. F., Harlow, M. K., & Meyer, D. R. (1950). Learning motivated by a manipulation drive. *Journal of Experimental Psychology*, 40, 228–234.
- Haruno, M., & Kawato, M. (2006). Heterarchical reinforcement-learning model for integration of multiple cortico-striatal loops: fMRI examination in stimulus-action-reward association learning. *Neural Networks*, 19, 1242–1254.
- Hayes-Roth, B., & Hayes-Roth, F. (1979). A cognitive model of planning. *Cognitive Science*, 3, 275–310.
- Hengst, B. (2002). Discovering hierarchy in reinforcement learning with HEXQ. *Proceedings of the International Conference on Machine Learning*, 19, 243–250.
- Holroyd, C. B., & Coles, M. G. H. (2002). The neural basis of human error processing: Reinforcement learning, dopamine, and the error-related negativity. *Psychological Review*, 109(4), 679–709.
- Hoshi, E., Shima, K., & Tanji, J. (1998). Task-dependent selectivity of movement-related neuronal activity in the primate prefrontal cortex. *Journal of Neurophysiology*, 80, 3392–3397.
- Houk, J. C., Adams, C. M., & Barto, A. G. (1995). A model of how the basal ganglia generate and use neural signals that predict reinforcement. In J. C. Houk & D. G. Davis (Eds.), *Models of information processing in the basal ganglia* (pp. 249–270). Cambridge: MIT Press.
- Joel, D., Niv, Y., & Ruppin, E. (2002). Actor-critic models of the basal ganglia: New anatomical and computational perspectives. *Neural Networks*, 15, 535–547.
- Johnston, K., & Everling, S. (2006). Neural activity in monkey prefrontal cortex is modulated by task context and behavioral instruction during delayed-match-to-sample and conditional prosaccade-antisaccade tasks. *Journal of Cognitive Neuroscience*, 18, 749–765.
- Jonsson, A., & Barto, A. (2001). *Automated state abstraction for options using the U-tree algorithm. Advances in neural information processing systems* (Vol. 13). Cambridge, MA: MIT Press (pp. 1054–1060).
- Jonsson, A., & Barto, A. (2005). A causal approach to hierarchical decomposition of factored MDPs. *Proceedings of the International Conference on Machine Learning*, 22.
- Kambhampati, S., Mali, A. D., & Srivastava, B. (1998). *Hybrid planning for partially hierarchical domains. Proceedings of the 15th national conference on artificial intelligence (AAAI-98)*. Madison, Wisconsin: AAAI Press.
- Kaplan, F., & Oudeyer, P.-Y. (2004). Maximizing learning progress: An internal reward system for development. In F. Iida, R. Pfeifer, & L. Steels (Eds.), *Embodied artificial intelligence*. Springer-Verlag.
- Kearns, M., & Singh, S. (2002). Near-optimal reinforcement learning in polynomial time. *Machine Learning*, 49, 209–232.
- Koechlin, E., Ody, C., & Kouneiher, F. (2003). The architecture of cognitive control in the human prefrontal cortex. *Science*, 302(5648), 1181–1185.
- Krueger, K. A., & Dayan, P. (2008). Flexible shaping. In *Cosyne (computational and systems neuroscience)*, Salt Lake City, Utah.
- Laird, J. E., Rosenbloom, P. S., & Newell, A. (1986). Chunking in soar: The anatomy of a general learning mechanism. *Machine Learning*, 1, 11–46.
- Landrum, E. R. (2005). Production of negative transfer in a problem-solving task. *Psychological Reports*, 97, 861–866.
- Lashley, K. S. (1951). The problem of serial order in behavior. In L. A. Jeffress (Ed.), *Cerebral mechanisms in behavior: The Hixon symposium* (pp. 112–136). New York, NY: Wiley.
- Lee, I. H., Seitz, A. R., & Assad, J. A. (2006). Activity of tonically active neurons in the monkey putamen during initiation and withholding of movement. *Journal of Neurophysiology*, 95, 2391–2403.
- Lee, F. J., & Taatgen, N. A. (2003). Production compilation: A simple mechanism to model complex skill acquisition. *Human Factors*, 45, 61–76.
- Lehman, J. F., Laird, J., & Rosenbloom, P. (1996). A gentle introduction to soar, an architecture for human cognition. In S. Sternberg & D. Scarborough (Eds.), *Invitation to cognitive science* (Vol. 4, pp. 212–249). Cambridge, MA: MIT Press.
- Li, L., & Walsh, T. J. (2006). Towards a unified theory of state abstraction for MDPs. *Paper presented at the 9th international symposium on artificial intelligence and mathematics*.
- Logan, G. D. (2003). Executive control of thought and action: In search of the wild homunculus. *Current Directions in Psychological Science*, 12, 45–48.
- Luchins, A. S. (1942). Mechanization in problem solving. *Psychological Monographs*, 248, 1–95.
- MacDonald, A. W., 3rd, Cohen, J. D., Stenger, V. A., & Carter, C. S. (2000). Dissociating the role of the dorsolateral prefrontal and anterior cingulate cortex in cognitive control. *Science*, 288(5472), 1835–1838.

- MacKay, D. G. (1987). *The organization of perception and action: A theory for language and other cognitive skills*. New York: Springer-Verlag.
- Mannor, S., Menache, I., Hoze, A., & Klein, U. (2004). Dynamic abstraction in reinforcement learning via clustering. In *Proceedings of the 21st international conference on machine learning* (pp. 560–567). ACM Press.
- Marthi, B., Russell, S. J., & Wolfe, J. (2007). Angelic semantics for high-level actions. *Paper presented at the 17th international conference on automated planning and scheduling (ICAPS 2007)*, Providence, RI.
- McGovern, A. (2002). *Autonomous discovery of temporal abstractions from interaction with an environment*. University of Massachusetts.
- Mehta, S., Ray, P., Tadepalli, P., & Dietterich, T. (2008). Automatic discovery and transfer of MAXQ hierarchies. In *International conference on machine learning*.
- Meltzoff, A. N. (1995). Understanding the intentions of others: Re-enactment of intended acts by 18-month-old children. *Developmental Psychology*, 31, 838–850.
- Menache, I., Mannor, S., & Shimkin, N. (2002). Dynamic discovery of subgoals in reinforcement learning. In *Proceedings of the 13th European conference on machine learning* (pp. 295–306).
- Middleton, F. A., & Strick, P. L. (2002). Basal-ganglia 'projections' to the prefrontal cortex of the primate. *Cerebral Cortex*, 12, 926–935.
- Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, 24, 167–202.
- Miller, G. A., Galanter, E., & Pribram, K. H. (1960). *Plans and the structure of behavior*. New York: Holt, Rinehart & Winston.
- Minton, S., Hayes, P. J., & Fain, J. (1985). Controlling search in flexible parsing. *Paper presented at the 9th international joint conference on artificial intelligence (IJCAI-85)*, Los Angeles.
- Miyamoto, H., Morimoto, J., Doya, K., & Kawato, M. (2004). Reinforcement learning with via-point representation. *Neural Networks*, 17, 299–305.
- Monsell, S. (2003). Task switching. *Trends in Cognitive Sciences*, 7(3), 134–140.
- Monsell, S., Yeung, N., & Azuma, R. (2000). Reconfiguration of task-set: Is it easier to switch to the weaker task? *Psychological Research*, 63, 250–264.
- Montague, P. R., Dayan, P., & Sejnowski, T. J. (1996). A framework for mesencephalic dopamine based on predictive Hebbian learning. *Journal of Neuroscience*, 16, 1936–1947.
- Morris, G., Arkadir, D., Nevat, A., Vaadia, E., & Bergman, H. (2004). Coincident but distinct messages of midbrain dopamine and striatal tonically active neurons. *Neuron*, 43, 133–143.
- Muhammad, R., Wallis, J. D., & Miller, E. K. (2006). A comparison of abstract rules in the prefrontal cortex, premotor cortex, inferior temporal cortex, and striatum. *Journal of Cognitive Neuroscience*, 18, 974–989.
- Nason, S., & Laird, J. E. (2005). Soar-RL: Integrating reinforcement learning with soar. *Cognitive Systems Research*, 6, 51–59.
- Newell, A., & Simon, H. A. (1963). GPS, a program that simulates human thought. In E. A. Feigenbaum & J. Feldman (Eds.), *Computers and thought* (pp. 279–293). New York: McGraw-Hill.
- Newtonson, D. (1976). Foundations of attribution: The perception of ongoing behavior. In J. H. Harvey, W. J. Ickes, & R. F. Kidd (Eds.), *New directions in attribution research* (pp. 223–248). Hillsdale, NJ: Erlbaum.
- O'Doherty, J., Critchley, H., Deichmann, R., & Dolan, R. J. (2003). Dissociating valence of outcome from behavioral control in human orbital and ventral prefrontal cortices. *Journal of Neuroscience*, 23, 7931–7939.
- O'Doherty, J., Dayan, P., Schultz, P., Deichmann, J., Friston, K., & Dolan, R. J. (2004). Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science*, 304(452–454).
- O'Reilly, R. C., & Frank, M. J. (2006). Making working memory work: A computational model of learning in prefrontal cortex and basal ganglia. *Neural Computation*, 18, 283–328.
- Oudeyer, P.-Y., Kaplan, F., & Hafner, V. (2007). Intrinsic motivation systems for autonomous development. *IEEE Transactions on Evolutionary Computation*, 11, 265–286.
- Parent, A., & Hazrati, L. N. (1995). Functional anatomy of the basal ganglia. I. The cortico-basal ganglia-thalamo-cortical loop. *Brain Research Reviews*, 20, 91–127.
- Parr, R., & Russell, S. (1998). Reinforcement learning with hierarchies of machines. *Advances in Neural Information Processing Systems*, 10, 1043–1049.
- Pashler, H. (1994). Dual-task interference in simple tasks: Data and theory. *Psychological Bulletin*, 116, 220–244.
- Petrides, M. (1995). Impairments on nonspatial self-ordered and externally ordered working memory tasks after lesions to the mid-dorsal part of the lateral frontal cortex in the monkey. *Journal of Neuroscience*, 15, 359–375.
- Piaget, J. (1936/1952). *The origins of intelligence in children* (M. Cook, Trans.). New York: International Universities Press (originally published, 1936).
- Pickett, M., & Barto, A. G. (2002). PolicyBlocks: An algorithm for creating useful macro-actions in reinforcement learning. In C. Sammut & A. Hoffmann (Eds.), *Machine learning: Proceedings of the 19th international conference on machine learning* (pp. 506–513). San Francisco: Morgan Kaufmann.
- Postle, B. R. (2006). Working memory as an emergent property of the mind and brain. *Neuroscience*, 139, 23–28.
- Rayman, W. E. (1982). Negative transfer: A threat to flying safety. *Aviation, Space and Environmental Medicine*, 53, 1224–1226.
- Ravel, S., Sardo, P., Legallet, E., & Apicella, P. (2006). Influence of spatial information on responses of tonically active neurons in the monkey striatum. *Journal of Neurophysiology*, 95, 2975–2986.
- Reason, J. T. (1992). *Human error*. Cambridge, England: Cambridge University Press.
- Redgrave, P., & Gurney, K. (2006). The short-latency dopamine signal: A role in discovering novel actions? *Nature Reviews Neuroscience*, 7, 967–975.
- Roesch, M. R., Taylor, A. R., & Schoenbaum, G. (2006). Encoding of time-discounted rewards in orbitofrontal cortex is independent of value. *Neuron*, 51, 509–520.
- Rolls, E. T. (2004). The functions of the orbitofrontal cortex. *Brain and Cognition*, 55, 11–29.
- Rougier, N. P., Noell, D. C., Braver, T. S., Cohen, J. D., & O'Reilly, R. C. (2005). Prefrontal cortex and flexible cognitive control: Rules without symbols. *Proceedings of the National Academy of Sciences*, 102, 7338–7343.
- Ruh, N. (2007). *Acquisition and control of sequential routine activities: Modelling and empirical studies*. University of London.
- Rumelhart, D., & Norman, D. A. (1982). Simulating a skilled typist: A study of skilled cognitive-motor performance. *Cognitive Science*, 6, 1–36.
- Rushworth, M. F. S., Walton, M. E., Kennerley, S. W., & Bannerman, D. M. (2004). Action sets and decisions in the medial frontal cortex. *Trends in Cognitive Sciences*, 8, 410–417.
- Ryan, R. M., & Deci, E. L. (2000). Intrinsic and extrinsic motivation. *Contemporary Educational Psychology*, 25, 54–67.
- Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, 13, 1926–1928.
- Saffran, J. R., & Wilson, D. P. (2003). From syllables to syntax: Multilevel statistical learning by 12-month-old infants. *Infancy*, 4, 273–284.
- Salinas, E. (2004). Fast remapping of sensory stimuli onto motor actions on the basis of contextual modulation. *Journal of Neuroscience*, 24, 1113–1118.
- Schank, R. C., & Abelson, R. P. (1977). *Scripts, plans, goals and understanding*. Hillsdale, NJ: Erlbaum.
- Schmidhuber, J. (1991). A possibility for implementing curiosity and boredom in model-building neural controllers. In *From animals to animats: Proceedings of the 1st international conference on simulation of adaptive behavior* (pp. 222–227). Cambridge: MIT Press.
- Schneider, D. W., & Logan, G. D. (2006). Hierarchical control of cognitive processes: Switching tasks in sequences. *Journal of Experimental Psychology: General*, 135, 623–640.
- Schoenbaum, G., Chiba, A. A., & Gallagher, M. (1999). Neural encoding in orbitofrontal cortex and basolateral amygdala during olfactory discrimination learning. *Journal of Neuroscience*, 19, 1876–1884.
- Schultz, W., Apicella, P., & Ljungberg, T. (1993). Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task. *Journal of Neuroscience*, 13, 900–913.
- Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, 275, 1593–1599.
- Schultz, W., Tremblay, K. L., & Hollerman, J. R. (2000). Reward processing in primate orbitofrontal cortex and basal ganglia. *Cerebral Cortex*, 10, 272–283.
- Shallice, T., & Burgess, P. W. (1991). Deficits in strategy application following frontal lobe damage in man. *Brain*, 114, 727–741.
- Shima, K., Isoda, M., Mushiake, H., & Tanji, J. (2007). Categorization of behavioural sequences in the prefrontal cortex. *Nature*, 445, 315–318.
- Shima, K., & Tanji, J. (2000). Neuronal activity in the supplementary and presupplementary motor areas for temporal organization of multiple movements. *Journal of Neurophysiology*, 84, 2148–2160.
- Shimamura, A. P. (2000). The role of the prefrontal cortex in dynamic filtering. *Psychobiology*, 28, 207–218.
- Simsek, O., Wolfe, A., & Barto, A. (2005). Identifying useful subgoals in reinforcement learning by local graph partitioning. In *Proceedings of the 22nd international conference on machine learning (ICML 05)*.

- Singh, S., Barto, A. G., & Chentanez, N. (2005). Intrinsically motivated reinforcement learning. In L. K. Saul, Y. Weiss, & L. Bottou (Eds.), *Advances in neural information processing systems 17: Proceedings of the 2004 conference* (pp. 1281–1288). Cambridge: MIT Press.
- Sirigu, A., Zalla, T., Pillon, B., Dubois, B., Grafman, J., & Agid, Y. (1995). Selective impairments in managerial knowledge in patients with prefrontal cortex lesions. *Cortex*, *31*, 301–316.
- Sommerville, J., & Woodward, A. L. (2005a). Pulling out the intentional structure of action: The relation between action processing and action production in infancy. *Cognition*, *95*, 1–30.
- Sommerville, J. A., & Woodward, A. L. (2005b). Infants' sensitivity to the causal features of means–end support sequences in action and perception. *Infancy*, *8*, 119–145.
- Suri, R. E., Bargas, J., & Arbib, M. A. (2001). Modeling functions of striatal dopamine modulation in learning and planning. *Neuroscience*, *103*, 65–85.
- Sutton, R. S., & Barto, A. G. (1990). Time-derivative models of pavlovian reinforcement. In M. Gabriel & J. Moore (Eds.), *Learning and computational neuroscience: Foundations of adaptive networks* (pp. 497–537). Cambridge: MIT Press.
- Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge, MA: 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*, 181–211.
- Tenenbaum, J. B., & Saxe, R. R. (Eds.). (2006). *Bayesian models of action understanding*. Cambridge, MA: MIT Press.
- Thrun, S. B., & Scwhartz, A. (1995). Finding structure in reinforcement learning. In G. Tesauro, D. S. Touretzky, & T. Leen (Eds.), *Advances in neural information processing systems: Proceedings of the 1994 conference*. Cambridge, MA: MIT Press.
- Wallis, J. D., Anderson, K. C., & Miller, E. K. (2001). Single neurons in prefrontal cortex encode abstract rules. *Nature*, *411*, 953–956.
- Wallis, J. D., & Miller, E. K. (2003). From rule to response: Neuronal processes in the premotor and prefrontal cortex. *Journal of Neurophysiology*, *90*, 1790–1806.
- Ward, G., & Allport, A. (1997). Planning and problem-solving using the five-disc Tower of London task. *Quarterly Journal of Experimental Psychology*, *50A*, 59–78.
- White, R. W. (1959). Motivation reconsidered: The concept of competence. *Psychological Review*, *66*, 297–333.
- White, I. M. (1999). Rule-dependent neuronal activity in the prefrontal cortex. *Experimental Brain Research*, *126*, 315–335.
- Wickens, J., Kotter, R., & Houk, J. C. (1995). Cellular models of reinforcement. In J. L. Davis & D. G. Beiser (Eds.), *Models of information processing in the basal ganglia* (pp. 187–214). Cambridge: MIT Press.
- Wolpert, D., & Flanagan, J. (2001). Motor prediction. *Current Biology*, *18*, R729–R732.
- Wood, J. N., & Grafman, J. (2003). Human prefrontal cortex: Processing and representational perspectives. *Nature Reviews Neuroscience*, *4*, 139–147.
- Woodward, A. L., Sommerville, J. A., & Guajardo, J. J. (2001). How infants make sense of intentional action. In B. F. Malle, L. J. Moses, & D. A. Baldwin (Eds.), *Intentions and intentionality: Foundations of social cognition*. Cambridge, MA: MIT Press.
- Yamada, S., & Tsuji, S. (1989). Selective learning of macro-operators with perfect causality. *Paper presented at the IJCAI-89, Detroit, MI*.
- Yan, Z., & Fischer, K. (2002). Always under construction: Dynamic variations in adult cognitive microdevelopment. *Human Development*, *45*, 141–160.
- Zacks, J. M., Braver, T. S., Sheridan, M. A., Donaldson, D. I., Snyder, A. Z., Ollinger, J. M., et al (2001). Human brain activity time-locked to perceptual event boundaries. *Nature Neuroscience*, *4*, 651–655.
- Zacks, J. M., Speer, N. K., Swallow, K. M., Braver, T. S., & Reynolds, J. R. (2007). Event perception: A mind/brain perspective. *Psychological Bulletin*, *133*, 273–293.
- Zacks, J. M., & Tversky, B. (2001). Event structure in perception and conception. *Psychological Bulletin*, *127*, 3–21.
- Zalla, T., Pradat-Diehl, P., & Sirigu, A. (2003). Perception of action boundaries in patients with frontal lobe damage. *Neuropsychologia*, *41*, 1619–1627.
- Zhou, W., & Coggins, R. (2002). Computational models of the amygdala and the orbitofrontal cortex: A hierarchical reinforcement learning system for robotic control. In R. I. McKay, & J. Slaney (Eds.), *Lecture notes AI: LNAI 2557* (pp. 419–430).
- Zhou, W., & Coggins, R. (2004). Biologically inspired reinforcement learning: Reward-based decomposition for multi-goal environments. In A. J. Ijspeert, M. Murata, & N. Wakamiya (Eds.), *Biologically inspired approaches to advanced information technology* (pp. 80–94). Berlin: Springer-Verlag.